Int J of Sci and Math Educ (2026) 24:16
https://doi.org/10.1007/s10763-025-10637-9

A Reliability Generalization Meta-Analysis of the Science Motivation Questionnaire II

Dilek Karaca[1] · Bahadır Namdar[2] · Fulden Güler Nalbantoğlu[2] · Begümay Tümer[2] · Burak Aydin[2,3]

Received: 6 May 2025 / Accepted: 31 October 2025 / Published online: 3 February 2026

Abstract
Science motivation is among the most important constructs affecting students' science learning and scientific thinking skills. One commonly utilized scale to measure students' science motivation is the Science Motivation Questionnaire II (SMQ-II), a 25-item, Likert-type, self-report scale. This study aimed to conduct a reliability generalization meta-analysis of the scale considering the Reliability Generalization Meta Analysis (REGEMA) guideline. Studies included in the analysis implemented the instrument between 2011–2024 and reported a Cronbach alpha value. Reliability evidence from 49 studies reporting a coefficient for the total score or at least for one of the five subscale scores was analyzed using a random effects model and Bonett's transformation. The pooled Cronbach alpha reliability coefficient was 0.94 for the total score and ranged from 0.83–0.88 for the subscale scores. Moderator analyses showed generally similar reliability estimates across studies, with different study types, languages, sample types, or sample sizes for the subscales. However, the test version, sample type, sample size, and female representation in the sample showed differences in reliability estimates for the total score. The empirical evidence from the SMQ-II's first 13 years reports high internal consistency across the scale's scores.

Keywords Meta-analysis · Reliability generalization · Science education · Science motivation

✉ Burak Aydin
burak.aydin@leuphana.de

[1] Project Coordination, Application and Research Center, Manisa Celal Bayar University, Manisa, Türkiye

[2] Faculty of Education, Ege University, Izmir, Türkiye

[3] Institute of Educational Sciences, Leuphana University, Lüneburg, Germany

Background

One of the fundamental goals of science education is to cultivate scientifically literate citizens who have a solid understanding of scientific knowledge and can use evidence-based reasoning to make informed decisions about issues related to science and society (Organization for Economic Co-operation and Development [OECD], 2007; Roberts, 2007). However, recent surveys indicate that students of varying ages lack basic scientific literacy (Aguilera & Perales-Palacios, 2020). Researchers have highlighted a bidirectional relationship between the enhancement of scientific literacy and science motivation (Van Vo & Csapó, 2022).

Given the important role of motivation to learn science, several studies have explored how to examine the level of motivation of college, secondary, and elementary students (Breland et al., 2023; Huda & Rohaeti, 2023; Jian-Xin et al., 2023; Meulenbroeks et al., 2024; Shin et al., 2023). Thus, researchers have utilized the Science Motivation Questionnaire-II (SMQ-II) developed by Glynn et al. (2011) for different purposes. For example, the SMQ-II has been employed in conjunction with other scales to develop a model and identify factors for Science, Technology, Engineering, and Mathematics (STEM) career interest formation among students (Alexopoulos et al., 2021; Lei, 2024; Riccitelli, 2015), as well as to evaluate undergraduate students' perceived motivation and engagement in game theme-based learning (Bónus et al., 2024; Moro & Billote, 2023; Watt, 2020).

Other studies have employed the SMQ-II to investigate motivation to learn specific science disciplines, such as chemistry and physics (Ardura & Pérez-Bitrián, 2018; Chen et al., 2023; de Souza et al., 2022; Hibbard et al., 2016; Kahraman, 2023; Kwarikunda et al., 2021; Lee & Mun, 2023; Lei, 2024; Salta & Koulougliotis, 2015, 2020; Schönfelder & Bogner, 2020). The instrument has been translated and adapted to different languages and cultural contexts to apply to Chinese (Dong et al., 2020), Spanish (Ardura & Pérez-Bitrián, 2018), Greek (Salta & Koulougliotis, 2020), German (Schumm & Bogner, 2016), Korean (Rachmatullah & Ha, 2019), and Turkish (Tosun, 2013) students. Researchers have used the SMQ-II to assess students' motivation in a fourth-year technical aircraft systems course (Ng & Chu, 2021), measure dental students' oral medicine learning motivation in Indonesia, and investigate the learning motivation profile of Japanese pharmacy students (Rahmayanti et al., 2020). The scale was also used to determine how different motivations of learners relate to their engagement in a massive open online astronomy course on the Coursera platform (Formanek et al., 2019). Consequently, discipline-specific versions of the SMQ-II emerged as researchers adapted the instrument for their studies. For instance, the Chemistry Motivation Questionnaire II (CMQ-II) (Austin et al., 2018) and Physics Motivation Questionnaire II (PMQ-II) (Kwarikunda et al., 2020) are frequently used versions of the original questionnaire that replace the word "science" with "chemistry" or "physics," respectively.

Overview of the Science Motivation Questionnaire (SMQ-II)

The SMQ-II (Glynn et al., 2011) represents a revised version of an instrument presented in an earlier study by Glynn et al. (2009). The first version of the scale consisted of 30 items in five subscales: intrinsic motivation and personal relevance (ten items), self-efficacy and assessment anxiety (nine items), self-determination (four items), career motivation (two items), and grade motivation (five items). Two years later, the researchers modified the questionnaire to improve the construct validity, based on the results of the first version of the questionnaire, focus group interviews, expert opinions, and a pilot study. Sixteen items remained unchanged in the revised version, and nine were added. As a result, the SMQ-II utilized in this meta-analysis consists of 25 items with five subscales: intrinsic motivation (IM), self-determination (SD), self-efficacy (SE), career motivation (CM), and grade motivation (GM). Five items in each subscale are scaled on a five-point Likert scale (from *never* to *always).* In the original validation study, reliability was reported as 0.92 for the total score, 0.92 for CM, 0.81 for GM, 0.89 for IM, 0.88 for SD, and 0.83 for SE.

During the revision of the SMQ-II scale, students' motivations were conceptualized more broadly in some respects and more narrowly in others. For example, intrinsic motivation was considered to include personal interest level, self-efficacy was considered to be related to evaluation anxiety, and extrinsic motivation was considered to be differentiated as grade motivation and career motivation. In this regard, it was stated that the SMQ-II scale targeted positive and mutually supportive motivation factors and took its final form by excluding negative motivation factors such as evaluation anxiety. It is also stated that the external motivation scale in the first version of the scale was examined in two separate sub-dimensions in SMQ-II: grade motivation and career motivation (Glynn et al., 2011).

Of the scale's subscales, intrinsic motivation refers to a natural interest, curiosity, and enjoyment of science learning, while career motivation refers to a belief in how learning science will help achieve future career goals. Self-determination refers to the control students believe they have over their own science learning processes and the effort they put into these processes, while self-efficacy represents students' beliefs/confidence that they can perform well and succeed in science. The grade motivation subscale assesses the importance of getting good grades in science courses and focusing on academic achievement (Glynn et al., 2011).

SMQ-II has been validated with undergraduate students (Glynn et al., 2009, 2011). Subsequent research has also focused on its use with middle school students (Ardura & Pérez-Bitrián, 2018; Razali, 2021; Salta & Koulougliotis, 2015; Toma et al., 2023), high school students (Conner, 2021; Dixon, 2017) and also preservice teachers (Haviz et al., 2024; Gomes de Freitas et al., 2020). The age-appropriate use of SMQ-II appears to be shaped by contextual factors, particularly the motivation for a specific subject (science, physics, chemistry, biology, etc.). This demonstrates that SMQ-II can be applied and adapted to different ages and different disciplines.

Reliability and Reliability Generalization (RG)

Reliability refers to the degree of consistency and stability of the scores from a measurement tool (Crocker & Algina, 1986). Errors in the measurement process reduce consistency; the source of error may vary from the scale itself, the sample (e.g., age, gender, motivation, etc.), the way the scale is applied, or the scoring complexity (Barnes et al., 2002; Dawis, 1987; Thompson & Vacha-Haase, 2000). It is critical to report reliability to support the reproducibility of research findings, particularly in the case of scores from a scale that is widely used in different languages and cultures, different sample groups, and different disciplines, such as the SMQ-II. The most appropriate option that allows a quantitative summary of reliability coefficients is the reliability generalization (RG) meta-analysis proposed by Vacha-Haase (1998). RG studies are meta-analyses that calculate an average reliability estimate and can explain the heterogeneity of the coefficients (Henson & Thompson, 2002; Rodriguez & Maeda, 2006; Sánchez-Meca et al., 2013; Vacha-Haase & Thompson, 2011). RG analysis also allows the examination of moderator variables that can explain a heterogeneity between the reported reliability coefficients (Sánchez-Meca et al., 2013). In other words, RG meta-analysis aims to estimate the average reliability of test scores based on various test applications in different samples and contexts, as well as evaluating the consistency across test applications. In addition, if differences in reliability coefficients are observed, it allows the determination of study and sample characteristics that may explain these differences from a statistical perspective (Thompson, 2002; Vacha-Haase, 1998).

The SMQ-II is a scale that exhibits strong psychometric properties and has become widely used internationally, after being adapted to more than ten languages to apply to various sample groups across cultures. Since the development of the SMQ-II, studies have been conducted on its construct validity (Glynn et al., 2011), measurement invariance for characteristics like gender and culture (Dong et al., 2020; Toma et al., 2023), construct validity in different languages and disciplines (Ardura & Pérez-Bitrián, 2018; de Souza et al., 2022; Toma et al., 2023) and adaptation (Salta & Koulougliotis, 2015). However, to our knowledge, no researchers have conducted an RG meta-analysis of the scale. Therefore, this study aims to conduct an RG meta-analysis on the SMQ-II scale due to its widespread use in different languages and cultures, its translations and adaptations into many languages, and its frequent application in the literature. The purpose of this study was to (a) determine a precise estimate of the overall reliability coefficient for the SMQ-II; (b) investigate how reliability coefficients vary; (c) determine whether there is heterogeneity between reliability estimates, and if so, conduct moderator analyses to identify study characteristics that could explain the variability.

Method

Selection Criteria

To improve its reporting, this paper follows the Reliability Generalization Meta Analysis (REGEMA, see supplementary), a 30-item checklist by Sánchez-Meca et al. (2021), which we provide as a supplement to this paper. First, a set of selection inclusion criteria was created to identify the studies for inclusion in the meta-analysis. Specifically we looked for studies that: (a) used the original SMQ-II (Glynn et al., 2011) or any of its adapted versions (Ardura & Pérez-Bitrián, 2018; Salta &

Koulougliotis, 2020), (b) were published or unpublished (e.g., Ph.D. dissertations) between 2011–2024, (c) were written in English, (d) allowed access to full text, (e) reported Cronbach's alpha reliability coefficient for the total SMQ-II or any of its subscales, and (f) focused on the target population in STEM fields regardless of age (e.g., university students, high school students, etc.). We also set exclusion criteria, specifically excluding studies that: (a) employed the earlier version of the SMQ (Glynn et al., 2009), (b) reported alpha values only from the original scale or from previous studies, (c) reported a reliability coefficient other than Cronbach's alpha, (d) were book chapters or conference proceedings, (e) added different items to total SMQ-II or any of its subscales, (f) focused on a field other than science education.

Search Strategies

Articles were obtained for the study sample by querying the Web of Science, ERIC, Scopus, and ProQuest databases with the search terms *Science Motivation Questionnaire II* and *SMQ II*. The search was limited to articles published in or after 2011, the when the SMQ-II scale was first used. We ensured search consistency across the research team by first searching for a certain number of articles together, before completing separate searches in different databases. After the random assignment of electronic databases to four authors, a folder was created for each database. The abstracts of the selected publications were read and evaluated, and if they met the criteria, the full text was examined. All related documents were collected in a main file and duplicated entries were easily detected and deleted.

Data Extraction and Procedure

A coding form was created following the RG meta-analysis literature as suggested by Henson and Thompson (2002), and the nature of the test was taken into consideration. We extracted Cronbach's alpha for the SMQ-II from the studies that reported at least one reliability estimate for the total score or subscale scores along with the following potential moderators: (a) study type (*article* or *dissertation/thesis)*; (b) language of the SMQ-II utilized (*English* or *other*); (c) study sample (*college students* or *other*); (d) test version (*original* or *adapted for a discipline*); (e) sample size; and (f) gender distribution (percentage of females).

The language of the SMQ-II utilized was coded into two categories. The studies in the dataset adapted the SMQ-II scale into twelve different languages; however, there was not enough data for each language category. Therefore, if the SMQ-II was delivered in English, the language moderator was coded as *English* or otherwise coded as *other*. Due to a similar lack of data for each category in the data set, the test version was also coded into two categories. If the scale was presented in its original form as developed by Glynn et al. (2009) where the word *science* was used, the test version moderator was coded as *original*; if it was adapted to another discipline (physics, chemistry, etc.), it was coded as *adapted*.

In some studies, more than one Cronbach alpha value was reported for more than one sample (experimental-control groups). If the reported values were calculated separately for different samples, they were coded separately. However, other studies reported more than one alpha value for the same sample (Ardura & Pérez-Bitrián, 2018; Salta & Koulougliotis, 2015). In these studies, although alpha values were calculated separately for different groups, there was also an alpha value for the combined sample where the groups were merged. In this case, only the alpha values for each group were included in the analysis. Some studies also added or removed items from the total or subdimensions of the scale (Park & Kim, 2021; Zhdanov et al., 2022). The valid item numbers and reliability values were coded in case of item removal. However, if new items were added to the scale, these data were not included in the analysis.

Searching for SMQ-II from 2011–2024 yielded a total of 259 studies, 76 of which were duplicate publications. Of the 183 studies included in the full-text review, 78 studies were excluded from the analysis because they did not meet the eligibility criteria: they utilized SMQ rather than SMQ-II, reported prior Cronbach's alpha values rather than computing, were not written in English, or conducted research in disciplines other than science (e.g., aviation, medicine). Thus, 49 studies that met the inclusion criteria were included in the analysis. The REGEMA flow chart in Fig. 1 presents the selection process details.

Reliability induction is a type of publication bias specific to RG meta-analysis, involving reporting a reliability coefficient from prior research but not for the study itself (Vacha-Haase et al., 2000). In our study, 30 of the 183 eligible studies did not report the reliability value, and nine reported the reliability value from the original SMQ-II study, resulting in a reliability induction rate of 21.31%. The reliability induction rate was obtained as an average of 78.6% in a systematic study of 100 RG meta-analyses, including more than 40,000 empirical studies in psychology (Sanchez-Meca et al., 2015). Hence, it can be stated that our study's reliability induction rate was at a reasonable level.

Intercoder Agreement

Three members of the research team coded the studies independently after a consensus meeting where three different studies were examined and discussed together. Then, the coders coded randomly selected ten studies that reported 60 reliability coefficients, to calculate Krippendorff's Alpha coefficient for intercoder agreement. This analysis was performed using the *irr* package (Gamer et al., 2019) in the R statistical software (R Core Team, 2025). The intercoder reliability coefficient for the three coders was 0.99, indicating near-perfect agreement.

Data Analysis

Similar to Sandoval-Lentisco et al. (2023), we conducted separate RG meta-analyses for each SMQ-II subscale before analyzing total scores. Since we assumed that there might be heterogeneity among population effect size estimates obtained from different studies, a random effects model was used in the analyses (Hedges & Vevea, 1998). All statistical analyses were conducted using the *metafor* package

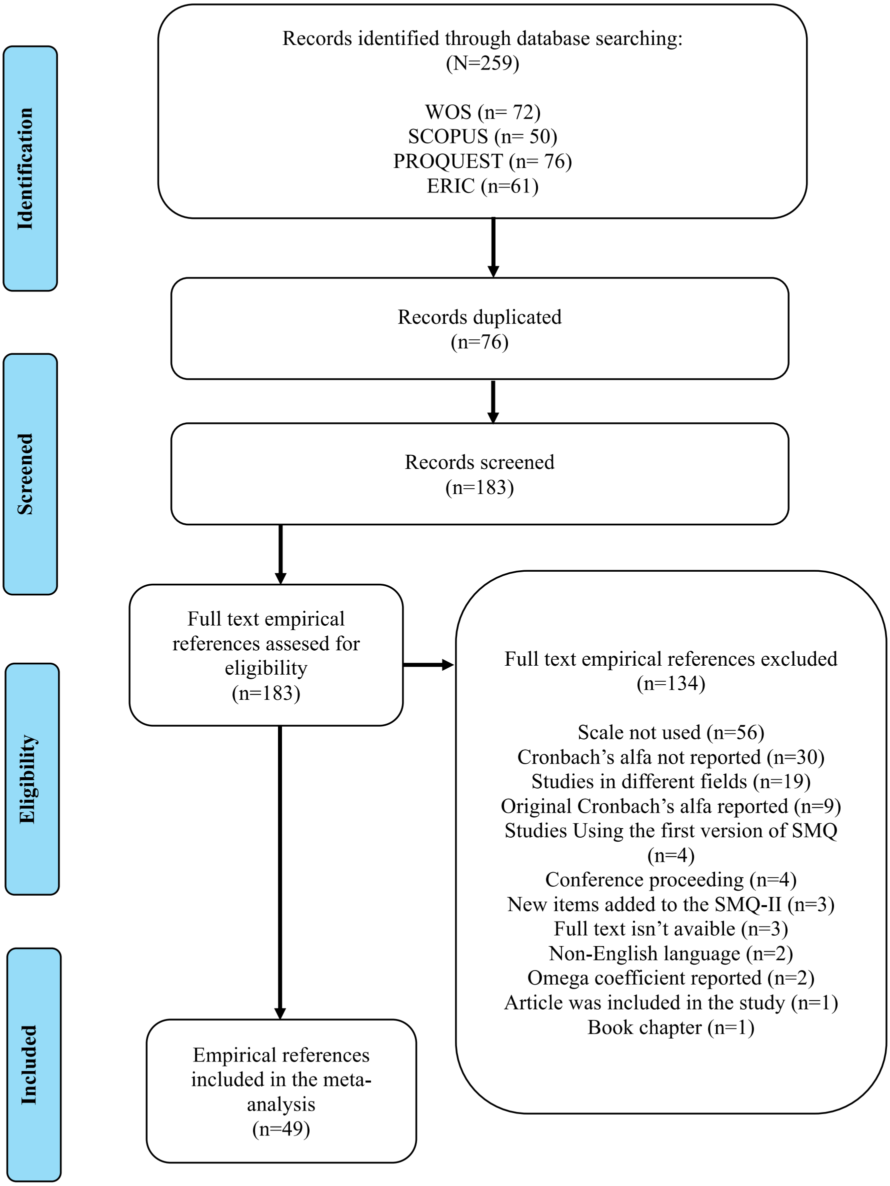

Fig. 1 REGEMA flow chart for SMQ-II RG meta-analysis

in R (Viechtbauer, 2010). Beretvas and Pastor (2003) suggest that alpha values reported in studies using the specified scale should not be used directly because they are usually negatively skewed. To normalize the distribution of reliability estimates and stabilize variances, it is recommended to apply one of Fisher's Z, Hakstian and Whalen's (1976) transformation or Bonett's (2002) transformation. Since Bonett's

(2002) transformation is recommended for reliability estimates that vary between 0 and 1, such as Cronbach's alpha or McDonald's omega (López-Ibáñez et al., 2024), Bonett's transformation was applied to the alpha values in this study. Therefore, we conducted the analyses by transforming the alpha values. We first computed outcome measures and sampling variances based on Bonett's transformation (2002) for the alpha values using $-\ln(1—\alpha_i)$, where α_i is the reported reliability coefficient. We then employed the random effects model with the restricted maximum likelihood (REML) and back-transformed the results using *transf=transf.iabt* argument to compute $1–1/e^{(estimate)}$ and reported point estimates and confidence intervals.

Various methods are suggested to determine heterogeneity between studies, such as the Q statistic, I^2 value or forest plot (Borenstein et al., 2021; Higgins et al., 2003; Huedo-Medina et al., 2006). The Q statistic is calculated by squaring the differences between the effect sizes in each study and the overall effect size obtained from the meta-analysis, multiplied by the weights of the studies. It is a value used to determine whether heterogeneity is present. The I^2 value, on the other hand, indicates the percentage of observed variance that is due to true heterogeneity (differences between studies) and provides the percentage of heterogeneity. Forest plots also enable the visual control of heterogeneity. In this study, we examined Q statistics, I^2 values, and forest plots to determine the heterogeneity between alpha coefficients. Statistically significant Q statistics (Cochran, 1954) and large I^2 values (> 75%) indicate heterogeneity (Higgins et al., 2003). To determine the reasonable reliability estimate interval of the studies, 95% confidence intervals were calculated, and forest plots containing these confidence intervals were created to provide a visual inspection of the effect sizes of the studies. For categorical and continuous moderator variables, heterogeneity analyses were performed using univariate meta-regression models in the *metafor* package (see supplementary for the data set and R syntax).

This RG meta-analysis was conducted with 49 studies: 40 articles (81.63%) and 9 theses (18.37%). The studies were conducted in over 20 countries, including the USA (the majority at 36.73%), Germany, Indonesia, Brazil, China, Greece, and Turkey. Thirty-six studies applied the SMQ-II scale in English (73.47%), while 13 studies applied it in eleven different languages (26.53%), including Portuguese, Greek, German, Chinese, Spanish, Russian, and Turkish. While 44.90% of the studies involved university students, 55.10% of the sample consisted of students at lower educational levels. The studies obtained data from 21,046 people in total. Since not all the included studies reported reliability coefficients for both the total score and for each subscale score, the number of reliability coefficients in each model varied, as explained in the results section.

Results

Publication Bias

We utilized funnel plots, Egger regression test (Egger et al., 1997), trim and fill method, and RG fail-safe N (Howell & Shields, 2008) analysis to examine publication bias for the coefficients of SMQ-II total scores. Egger's regression test showed a statistically significant asymmetry ($t=-3.530$, sd=24, $p=.002$), possibly due to a small study effect or possible publication bias. However, the trim and fill method analysis determined that there were no missing studies on the right side (see supplementary). The funnel plot in Fig. 2 shows a symmetrical distribution, but with some asymmetry due to points remaining outside in the upper right region. Following Howell and Shields (2008), we first computed the lower bound estimate of unweighted alpha for the total score in 26 included studies (see Table S4 in the supplementary). The average alpha for these 26 studies was.93 with a standard deviation of.03, with a plausible worst-case reliability coefficient of.80 standard deviations below the average. For the 134 eligible but not useful studies, average alpha was estimated as.906, hence resulting in a lower bound estimate of.91. Again, following Howell and Shields (2008) with similar criteria and a threshold reliability of.80, the RG fail-safe N was computed as 39. In other words, the mean population reliability of SMQ-II scores would be less than.80 if there were 39 file drawer studies with an unweighted average reliability of.712, a value far below the estimated lower bound. Overall, the findings obtained in the following RG meta-analysis are expected to be highly robust to publication bias.

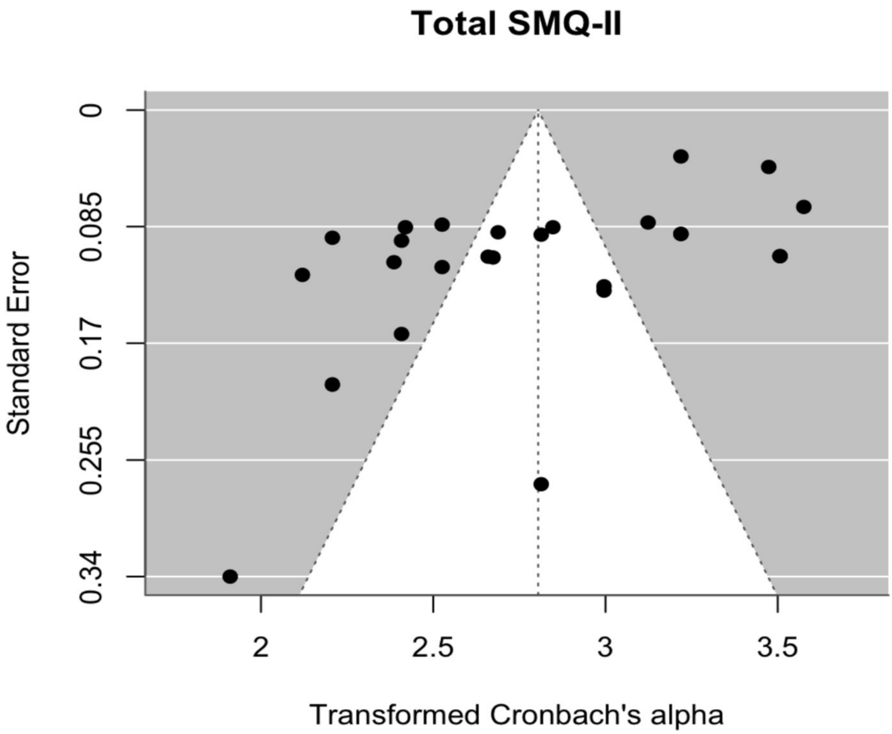

Fig. 2 Funnel plot of transformed Cronbach alpha coefficients for SMQ-II total score

Mean Reliability and Heterogeneity

Table 1 reports the year, type, and country for the included studies. Forty of the studies were articles, whereas nine were theses. Although the studies conducted in more than 20 countries, most were conducted in the USA (36.73%).

For the subscales, 47 alpha values were obtained from 32 different studies for CM, 45 from 33 studies for GM, 43 from 32 studies for IM, 46 from 31 studies for SD, and 50 from 34 studies for SE. For the SMQ-II total score, 27 values were obtained from 23 different studies. The data analyzed had been collected from 12,434 individuals for the CM subscale, 12,683 for GM, 16,211 for IM, 11,559 for SD, 13,675 for SE, and 6,881 individuals from more than three countries for the total scale (see supplementary). Figures 3, 4, 5, 6, 7 present forest plots of reliability coefficients for the overall scale and each subscale.

The total scale reliability values of the studies ranged from.85 (Moro & Billote, 2023) to.97 (Huda & Rohaeti, 2023; Lee & Mun, 2023; Razali et al., 2020). The forest plot in Fig. 8 shows that the weights of all studies are close to each other, apart from a few studies (Lei, 2024; Migalang & Azuelo, 2020; Watt, 2020).

Table 2 shows the main results for the total and each of the five subscales of the SMQ-II scale. The weighted average calculated for alpha (α_+) for the total score of the SMQ-II scale was 0.94 (95% CI[.93,.95]) and was statistically significant at the $p<.001$ level. The Q statistic was significant ($Q=664.0269$, $p<.001$) and the I^2 statistic was 96.36.

Among the subscales of the SMQ-II, $\alpha_+=0.88$ (95% CI:[0.86, 0.90]) for CM, $\alpha_+=0.83$ (95% CI:[0.80, 0.85]) for GM, $\alpha_+=0.85$ (95% CI:[0.82, 0. 87]), $\alpha_+=0.84$ (95% CI:[0.82, 0.86]) for SD, and $\alpha_+=0.85$ (95% CI:[0.84, 0.87]) for SE were statistically significant at $p<.001$ level. Cochran's Q statistic and I^2 statistics were found to be statistically significant in the CM ($Q=1436.1580$, $p<.0001$, $I^2=96.96$), GM ($Q=1960.1004$, $p<.0001$, $I^2=96.98$); IM ($Q=1412.\ 9436$, $p<.0001$, $I^2=97.40$), SD ($Q=1059.1763$, $p<.0001$, $I^2=95.88$), and SE subscales ($Q=1366.1194$, $p<.0001$, $I^2=95.95$). Accordingly, the statistically significant Q statistics and high I^2 values indicate a high level of heterogeneity among the studies reporting reliability values for the subscales of the SMQ-II scale.

Moderator Analysis

The Q statistics and high I^2 values (>92) show significant heterogeneity between the mean alpha coefficients for the SMQ-II total score and its subscales, as supported by the forest plots. Moderator analyses were conducted to investigate potential sources of heterogeneity, with study type, study language, sample type, and test version as categorical variables, and sample size and female percentage as continuous variables. While there were missing values for continuous moderator variables in the data set, no imputation was made for these values, and they were considered missing data.

Table 1 Studies included in the analysis

No	Study	Year	Type	Country	No	Study	Year	Type	Country
1	Shin et al. (2018)	2018	A	South Korea	26	Migalang and Azuelo (2020)	2020	A	Philippines
2	Chen et al. (2023)	2023	A	The USA	27	Haviz et al. (2024)	2024	A	Indonesia
3	Salta and Koulougliotis (2015)	2015	A	Greece	28	Breland et al. (2023)	2023	A	The USA
4	You et al. (2018)	2018	A	The USA	29	Gomes de Freitas et al. (2020)	2020	A	Brazil
5	Toma et al. (2023)	2023	A	Brazil	30	Hugerat et al. (2020)	2020	A	Israel
6	Bónus et al. (2024)	2024	A	Hungary	31	Chumbley et al. (2022)	2022	A	The USA
7	Salta and Koulougliotis (2020)	2020	A	Greece	32	Zhdanov et al. (2022)	2022	A	Russia
8	Hibbard et al. (2016)	2016	A	The USA	33	Kulesza et al. (2022)	2022	A	The USA
9	Razali (2021)	2021	A	Malaysia	34	Reece and Butler (2017)	2017	A	The USA
10	Schmid and Bogner (2017)	2017	A	Germany	35	Alexopoulos et al. (2021)	2021	A	Italy
11	Moro and Billote (2023)	2023	A	Philippines	36	Schönfelder & Bogner (2020)	2020	A	Germany
12	Schumm & Bogner (2016)	2016	A	Germany	37	Rachmatullah and Ha (2019)	2019	A	Indonesia
13	de Souza et al. (2022)	2022	A	Brazil	38	Iyer (2017)	2017	A	India
14	Willoughby et al. (2024)	2024	A	New Zealand	39	Rachmatullah et al. (2018)	2018	A	Multiple
15	Dong et al. (2020)	2020	A	China	40	Lee and Mun (2023)	2023	A	Korea
16	Austin et al. (2018)	2018	A	The USA	41	Bernard (2018)	2018	T	The USA
17	Formanek et al. (2019)	2019	A	Multiple	42	Lawson (2020)	2020	T	The USA
18	Dixon and Wendt (2021)	2021	A	The USA	43	Christiansen (2015)	2015	T	The USA
19	Kwarikunda et al. (2021)	2021	A	Uganda	44	Conner (2021)	2021	T	The USA
20	Razali et al. (2020)	2020	A	Malaysia	45	Dixon (2017)	2017	T	The USA
21	Ardura and Pérez-Bitrián (2018)	2018	A	Spain	46	Riccitelli (2015)	2015	T	The USA
22	Huda and Rohaeti (2023)	2023	A	Indonesia	47	Heim (2020)	2020	T	The USA
23	Kwarikunda et al. (2020)	2020	A	Uganda	48	Pope (2018)	2018	T	The USA
24	Kahraman (2023)	2023	A	Turkey	49	Watt (2020)	2020	T	The USA
25	Lei (2024)	2024	A	China					

A Article, *T* Thesis

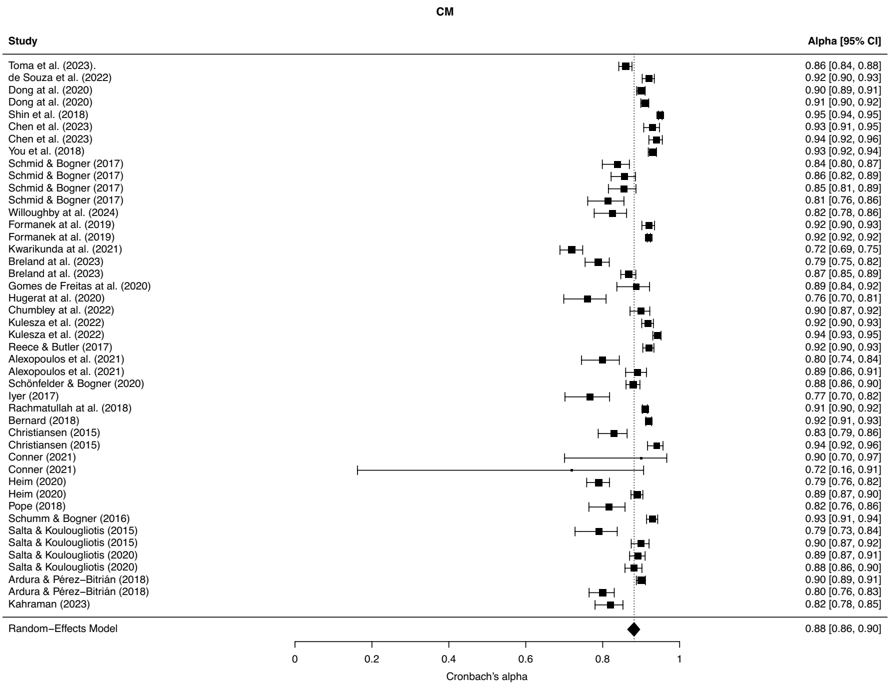

Fig. 3 Forest plot of the α_+ coefficients reported for the career motivation subscale

The results of the meta-regression analysis for categorical moderator variables are presented in Table 3. The results show that there are no significant differences in the reliability coefficients of moderator variables in the CM, GM, and IM subscales (p>.05). However, the test version was a statistically significant moderator variable in the SD (Q_M=7.712, p=.006) and SE (Q_M=9.726, p=.002) subscales. The results of the analysis revealed that the average reliability coefficients of the original test were higher than those of the adapted test. In addition, the sample type was a statistically significant moderator variable for the total scale (Q_M=10.540, p=.001). The average reliability value calculated for the samples consisting of university students (0.92) was lower than the samples consisting of other students (0.95). Table 3 presents the average reliability coefficients of the moderator variables for the total scale and subscales.

Table 4 presents the results of the meta-regression analysis of continuous moderator variables. The sample size variable had a statistically significant effect on the total scale (b_j=.001, SE=-.000, p=.028, R^2=13.87). Although this finding indicates that increasing the sample size significantly affects the reliability coefficient calculated for the total scale, it also indicates that the effect is relatively small. However, the sample size does not have a statistically significant effect on the reliability coefficient of any of the subscales (p>.05).

Gender distribution, one of the moderator variables, shows a statistically significant negative effect for GM (b_j=-.011, SE=.005, p=.028, R^2=12.08). This finding

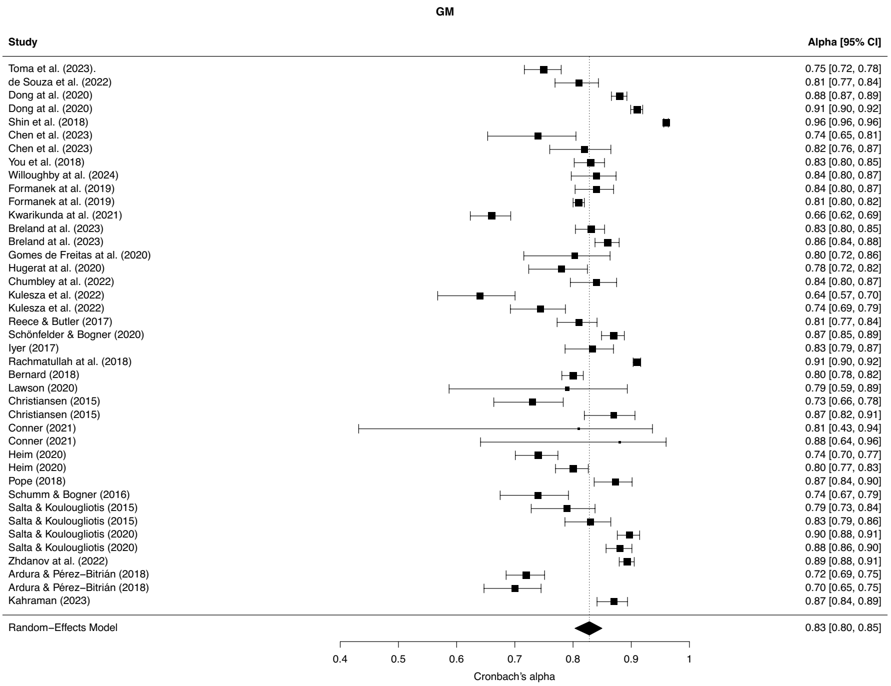

Fig. 4 Forest plot of the α_+ coefficients reported for the grade motivation subscale

indicates that an increase in the proportion of female participants has a significant, but relatively small effect on the reliability coefficient for the GM subscale. For the total scale and other subscales (CM, IM, SD, and SE), gender distribution had no significant effect ($p > .05$).

Conclusions and Discussion

This RG meta-analysis study aimed to estimate the mean reliability coefficient of the SMQ-II scale and its subscales and examine moderator variables that may cause variability between studies. For this purpose, a meta-analysis was conducted on 49 studies obtained within the scope of the criteria determined. The mean reliability coefficient for the SMQ-II total scale score was calculated as 0.94. This value is very close to the original reliability coefficient of 0.92 reported by Glynn et al. (2011), but still slightly higher. The mean reliability coefficients of the SMQ-II subscales were between 0.83–0.88, which can be expressed as "very good" according to DeVellis (2003). Furthermore, these values are above the acceptable lower limit of 0.70 proposed by Nunnally (1978) for measurement instruments and the widely accepted 0.80 in the social sciences. When compared with the original reliability coefficients of the scale, the mean reliability coefficient was higher in two subscales (GM and SE) and lower in three subscales (CM, IM, and SD). A large heterogeneity

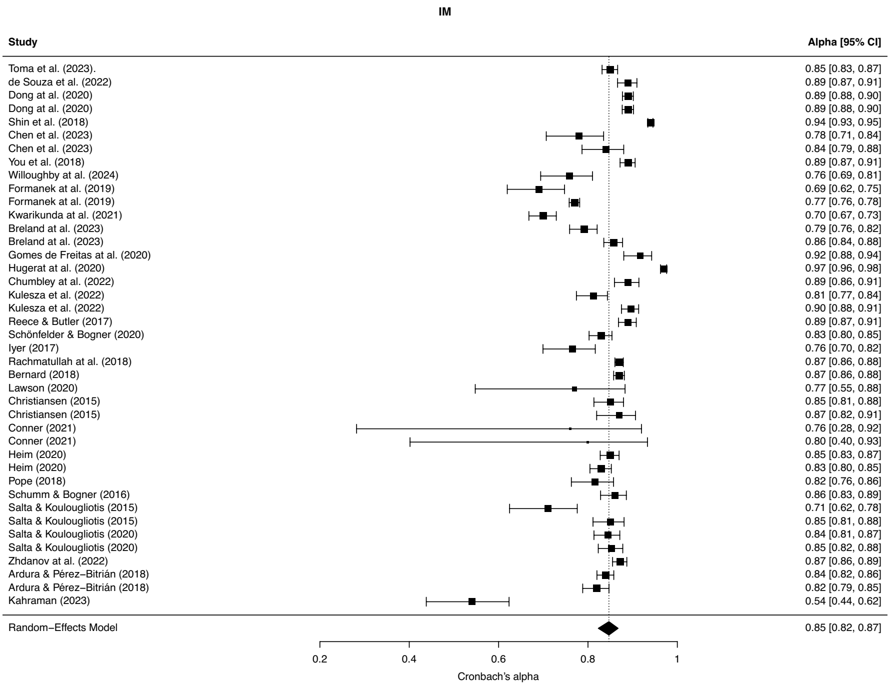

Fig. 5 Forest plot of the α_+ coefficients reported for the intrinsic motivation subscale

($I^2 = 96.36$) was found between the coefficients for the SMQ-II total scale. Thus, we recommend that researchers who plan to use the scale calculate the reliability value for their own samples and not make reliability induction.

Our results also showed that the reliability values for the total scale varied according to the type of sample (university versus lower-level students). The mean reliability value calculated for samples consisting of university students (0.92) was lower than that for samples consisting of lower-level students (0.95). While the scientific motivation of university students is related to individual interest and academic/career goals, it may depend on teacher or parental guidance for students at other educational levels. Second, students' levels of intellectual and cognitive development may affected their responses to the scale. Less advanced students may have perceived and made sense of abstract concepts such as scientific motivation differently than more advanced students. The result that the sample type affects the average reliability estimate mirrors the findings of previous studies by Vacha-Haase (1998), Yin and Fan (2000), Eser and Doğan (2023), and Batı and Irmak (2024).

Another important moderator variable for total score reliability was sample size. An increase in sample size significantly affected the reliability coefficient, but this effect was statistically expected, as large samples provide more stable results than small samples. The positive relationship between sample size and reliability coefficients has yielded similar results in other studies (Sandoval-Lentisco et al., 2023).

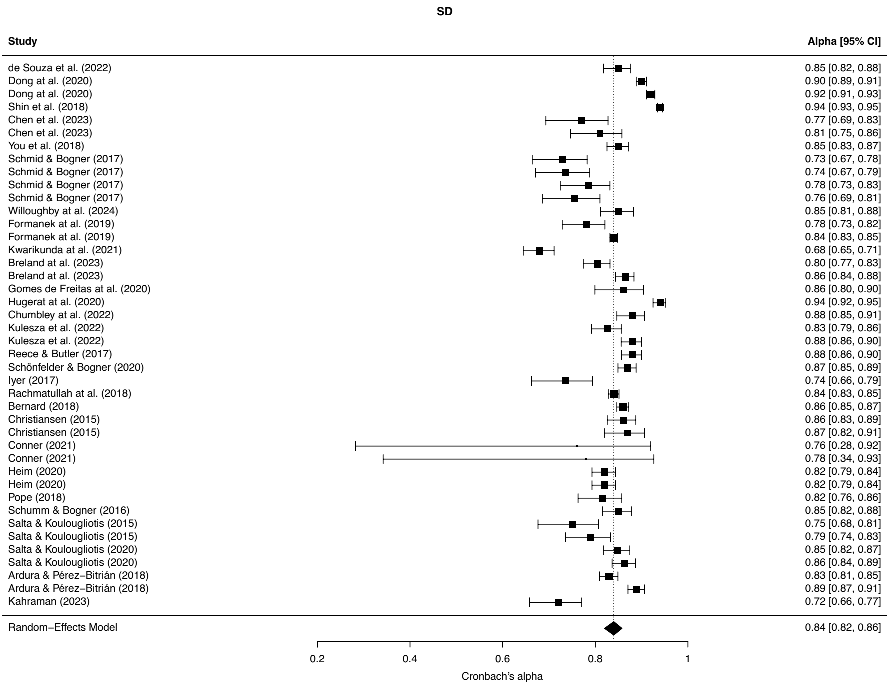

Fig. 6 Forest plot of the α_+ coefficients reported for the self-determination subscale

Significant heterogeneity was also found between the reliability coefficients in the subscales. The moderator analyses to gauge heterogeneity concluded that the type of study, language, sample size, and sample type variables did not affect the subscale reliability values; however, test version, one of the moderator variables, did impact the change of reliability values in the SD and SE subscales. Moreover, the average reliability coefficients of the original test version were higher than the adapted test version.

Another moderator variable, female percentage was found to have an effect on the reliability values in the GM subscale. However, this effect was relatively small. This result could stem from the differing levels of gender equality in the countries where the studies in the sample were conducted. Earlier measurement invariance studies for the SMQ-II did not reveal significant differences between genders (Dong et al., 2020; Toma et al., 2023). However, our results suggest that new studies should employ measurement invariance studies to assure that the SMQ-II measures similarly across genders.

The primary contribution of this reliability generalization is the identification of variability in reliability estimates across the subscales of the questionnaire. Such interpretations can lead researchers to interpret each of the subscales independently rather than assuming overall measurement quality. For example, consistently higher reliability scores across the subscales may warrant their use in research, while subscales with lower reliability may suggest cautious use of such subscales

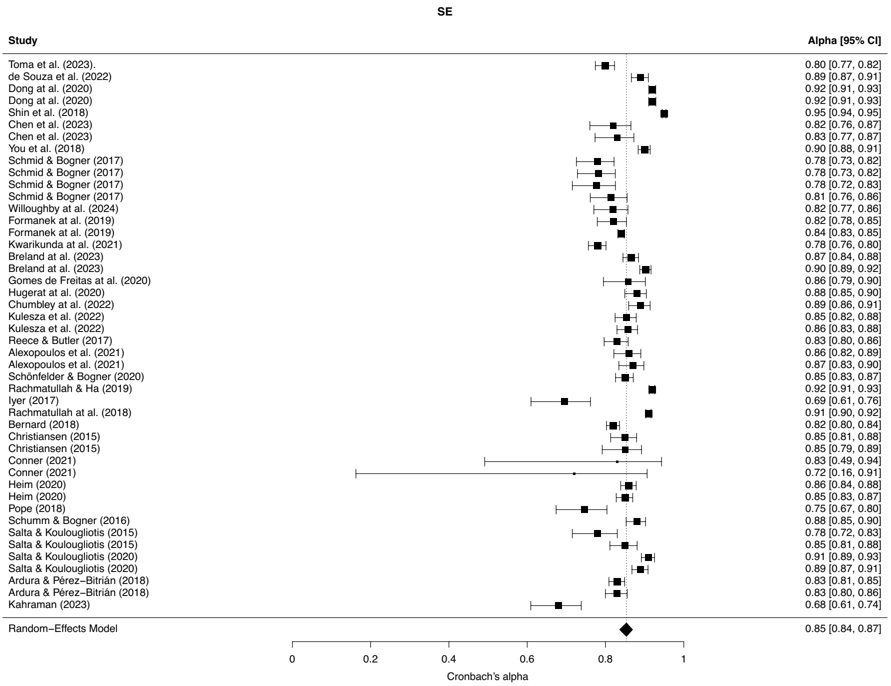

Fig. 7 Forest plot of the α_+ coefficients reported for the self-efficacy subscale

or instrument modification. Moreover, the moderators may also indicate the contextual factors that influence the scores obtained from the participants. These findings can influence the preferred use of SMQ-II and its subscales, while helping researchers to determine whether adaptations are necessary for a more valid and reliable measurement.

The present study's analysis of research conducted on individuals from more than 15 countries, with sample sizes ranging from 6,800 to 16,000, revealed very high reliability coefficients for both the total and subscale scores of the SMQ-II. However, reliability is not a fixed characteristic of scales, rather a characteristic of the scores obtained at the end of the measurements (Crocker & Algina, 1986). Organizations such as the APA (Wilkinson, 1999), American Educational Research Association, and National Council on Measurement in Education recommend that researchers calculate and report the reliability values of their own studies. The literature suggests that many researchers ignore this recommendation and present the scale by referring either to the original reliability coefficient or to the reliability coefficient obtained in another study, which Vacha-Haase et al. (2000) termed "reliability induction." Therefore, it is important for researchers to report the reliability coefficients obtained from their own studies, regardless of which scale they use. Investigations have shown that researchers' most commonly reported reliability coefficient is Cronbach's alpha (Scherer & Teo, 2020). This may be because Cronbach's Alpha is easy to calculate, does not require the scale to be reapplied, and can

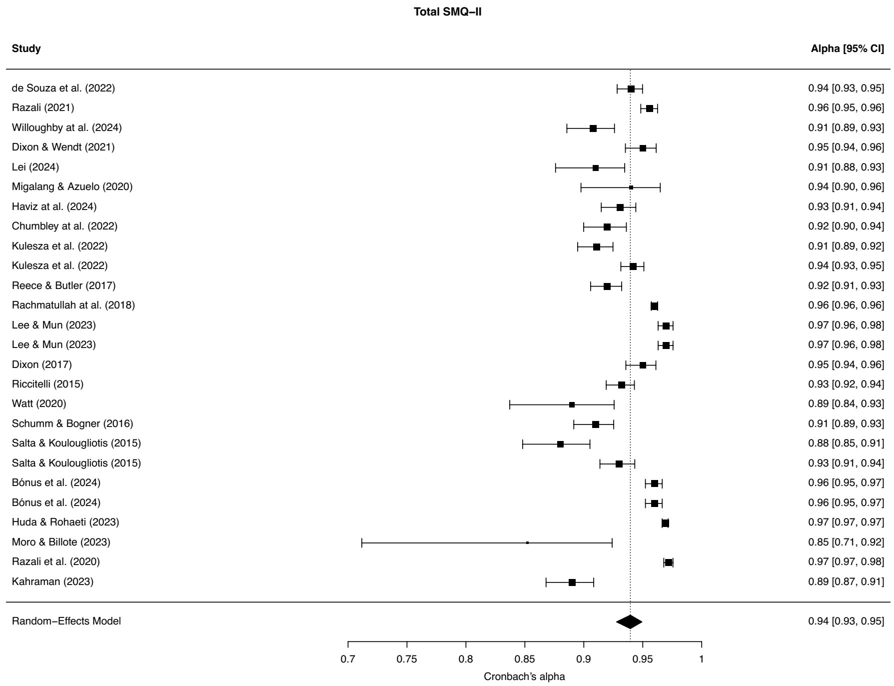

Fig. 8 Forest plot of the α_+ coefficients reported for total scale

Table 2 Mean reliability coefficients, 95% confidence and credibility/prediction intervals, and heterogeneity statistics for each subscale

			95% CI		95% PI			
	k	α_+	LL	UL	LL	UL	*Q*	I^2
CM	45	.88	.86	.90	.72	.95	1436.1580*	96.96
GM	41	.83	.80	.85	.60	.93	1960.1004*	96.98
IM	41	.85	.82	.87	.62	.94	1412.9436*	97.40
SD	42	.84	.82	.86	.67	.92	1059.1763*	95.88
SE	46	.85	.84	.87	.70	.93	1366.1194*	95.95
Total	26	.94	.93	.95	.85	.98	664.0269*	96.36

Abbreviations: *k*, number of studies; α_+, mean coefficients α; LL and UL, lower and upper limits of the 95% confidence and credibility/prediction intervals for α; *Q*, Cochran's heterogeneity *Q* statistic; I^2, heterogeneity index; *$p<.00$, *CM* career motivation, *GM* grade motivation, *IM* intrinsic motivation, *SD* self-determination, *SE* self-efficacy.

be calculated directly in many software programs. In addition, calculating other reliability coefficients may require more complex statistical knowledge and calculation skills. Another reason for this situation may be that researchers are not familiar with

Table 3 Results of categorical moderator analyses for total scale and subscales

				95%			
Moderator variable		*k*	α	LL	*UL*	Q_M	*p*
CM							
Study Type	Article	39	.88	.87	.90	.251	.617
	Thesis	8	.87	.82	.91		
Language	English	35	.88	.86	.90	.009	.925
	Other	12	.88	.85	.91		
Sample Type	University students	24	.89	.87	.91	1.527	.217
	Other	23	.87	.84	.89		
Test Version	Original	26	.88	.86	.90	.018	.892
	Adapted	21	.88	.86	.90		
GM							
Study Type	Article	36	.83	.80	.86	.384	.535
	Thesis	9	.81	.74	.86		
Language	English	32	.82	.79	.85	.182	.670
	Other	13	.84	.79	.87		
Sample Type	University students	27	.83	.79	.85	.090	.765
	Other	18	.83	.79	.87		
Test Version	Original	25	.84	.81	.87	2.228	.136
	Adapted	20	.81	.76	.84		
IM							
Study Type	Article	34	.85	.82	.87	.184	.668
	Thesis	9	.84	.77	.88		
Language	English	30	.85	.82	.87	.165	.685
	Other	13	.84	.79	.88		
Sample Type	University students	25	.84	.80	.86	1.433	.231
	Other	18	.86	.83	.89		
Test Version	Original	25	.86	.83	.88	2.873	.090
	Adapted	18	.82	.78	.86		
SD							
Study Type	Article	38	.84	.82	.86	.029	.865
	Thesis	8	.84	.78	.88		
Language	English	35	.84	.81	.86	.348	.555
	Other	11	.85	.81	.88		
Sample Type	University students	26	.84	.81	.86	.187	.666
	Other	20	.84	.81	.87		
Test Version	Original	23	.86	.84	.88	7.712	**.006*****
	Adapted	23	.81	.78	.84		
SE							
Study Type	Article	42	.86	.84	.87	1.275	.259
	Thesis	8	.83	.77	.87		
Language	English	38	.85	.83	.87	.388	.533
	Other	12	.86	.83	.89		

Table 3 (continued)

Moderator variable		k	α	95% LL	95% UL	Q_M	p
Sample Type	University students	26	.85	.82	.87	.897	.344
	Other	24	.86	.84	.88		
Test Version	Original	27	.87	.85	.89	9.726	**.002******
	Adapted	23	.82	.79	.85		
Total scale							
Study Type	Article	24	.94	.93	.95	.382	.536
	Thesis	3	.93	.88	.96		
Language	English	17	.94	.92	.95	.005	.945
	Other	10	.94	.92	.95		
Sample Type	University students	10	.92	.89	.93	10.540	**.001******
	Other	17	.95	.94	.96		
Test Version	Original	14	.95	.93	.96	1.864	.172
	Adapted	13	.93	.91	.95		

k = number of alpha coefficients; b_j = unstandardized regression coefficient; p = p value; Q_E = statistic to test for residual heterogeneity; SE = standard error; R^2 = proportion of variance accounted for by the predictor. ***$p < .001$

Table 4 Meta-regression results for continuous variables

Moderator variable		k	b_j	SE	p	R^2	Q_E
CM	Sample Size	45	.000	.000	.157	2.37	1229.637***
	Female Percent	40	-.002	.005	.781	.00	1164.501***
GM	Sample Size	41	.000	.000	.364	.00	1959.794***
	Female Percent	36	-.011	.005	**.028**	**12.08**	1342.208***
IM	Sample Size	41	-.000	.000	.900	.00	1236.701***
	Female Percent	36	-.005	.004	.249	1.87	945.370***
SD	Sample Size	42	.000	.000	.395	.00	1058.111***
	Female Percent	37	-.005	.004	.209	2.17	715.175***
SE	Sample Size	46	.000	.000	.112	2.94	1363.766***
	Female Percent	41	-.005	.004	.248	.99	1065.478***
Total scale	Sample Size	26	.001	.000	**.028**	**13.87**	486.755***
	Female Percent	14	-.016	.009	.088	14.43	281.182***

k = number of alpha coefficients; b_j = unstandardized regression coefficients; SE = standard error; R^2 = explained variance; Q_E = test statistic of the test for residual heterogeneity; *$p < .05$; ***$p < .001$, *CM* career motivation, *GM* grade motivation, *IM* intrinsic motivation, *SD* self-determination, *SE* self-efficacy

these methods, or that academic journal referees and editors are also familiar with Cronbach's Alpha. A review of the literature shows that reliability generalization in meta-analysis studies is predominantly based on Cronbach's Alpha, while studies using other reliability coefficients such as McDonald's Omega or KR-21 (Husain et al., 2025; Mack et al., 2024; Sandoval-Lentisco et al., 2023; Tsui et al., 2025). In this context, it is crucial for researchers to report different reliability coefficients rather than focusing solely on Cronbach's Alpha. The widespread use of different reliability coefficients will contribute to the formation of richer and more diverse data pools in reliability generalization meta-analysis studies, the comparability of measurement tools in different contexts, and a more comprehensive evaluation of the psychometric properties of measurement tools.

Implications and Limitations

In our RG meta-analysis of the SMQ-II, the type of study, study language, sample type, test version, sample size, and female percentage were identified as moderator variables to determine the sources of heterogeneity. Future studies can focus on different moderator variables to investigate the sources of heterogeneity. More in-depth analyses are needed, especially for sample types that show statistically significant differences across sources of heterogeneity. Future research could focus on systematically examining how reliability coefficients vary across different age groups, educational levels, and cultural contexts. Such studies would make important contributions to assessing the applicability of the scales across different demographic and socio-cultural groups.

The SMQ-II has been translated into many languages and modified for use in a variety of fields, including biology, chemistry, physics, and aviation. However, our study only considered English-language research, which poses another limitation since it results in the exclusion of research in other languages. At the same time, although there are adaptation studies in many disciplines, the test version variable of the scale could not be examined separately according to these disciplines due to the limited number of studies in the data set. In our study, the studies using the original form of the scale were coded as "original" in the test version, and the versions adapted to different disciplines were coded as "adapted" and included in the analysis. In this context, the inability to conduct detailed analyses on a discipline-by-discipline basis regarding the translation and adaptation processes of the scale also creates a limitation. Since we only reviewed studies that applied the SMQ-II to science education, future RG meta-analyses of the scale should include studies in different disciplines (e.g., aviation or medicine) and languages. Detailed reporting of different reliability coefficients, sample characteristics, and descriptive statistics of the sampled studies would allow future research to focus on new coefficients and moderator variables.

This study is limited to examining the reliability generalization of the SMQ-II scale and its subscales. Future research could increase the scope and generalizability of the findings by enabling a comparative examination of the long version of the SMQ scale or different motivation scales measuring similar constructs (e.g., another scale related to science motivation).

Technological advances and digital learning environments affect student motivation towards science and may consequently change how students respond to tests and questionnaires. In this context, differences can be observed between the reliability coefficients obtained from SMQ-II assessments applied in traditional paper-and-pencil form and the coefficients obtained from digital assessments. However, there is not enough data in the studies examined in the present research to evaluate this difference. We recommend that this situation be systematically addressed in future RG meta-analysis studies. Despite the aforementioned limitations, it is still reasonable to assert that the SMQ-II scale has strong internal consistency and robustness to adaptations for related disciplines.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10763-025-10637-9.

Author Contributions All authors contributed to the study, specifically:

DK Conceptualization, methodology, software, formal analysis, data curation, writing & editing, visualization.

BN Conceptualization, collection and interpretation of data, writing & editing, supervision.

FGN Conceptualization, collection and interpretation of data, writing.

BT Collection and interpretation of data, writing.

BA Conceptualization, methodology, software, formal analysis, data curation, writing & editing, visualization, supervision.

Funding Open Access funding enabled and organized by Projekt DEAL. This publication was funded by the Open Access Publication Fund of the Leuphana University Lüneburg.

Data Availability See supplementary files.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication All Authors have given consent to publish this manuscript.

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

References with an asterisk (*) indicate studies included in the reliability generalization meta-analysis in this study.

Aguilera, D., & Perales-Palacios, F. (2020). What effects do didactic interventions have on students' attitudes towards science? A meta-analysis. *Research in Science Education, 50*, 573–597. https://doi.org/10.1007/s11165-018-9702-2

*Alexopoulos, A. N., Paolucci, P., Sotiriou, S. A., Bogner, F. X., Dorigo, T., Fedi, M., Menasce, D., Michelotto, M., Paoletti, S., & Scianitti, F. (2021). The colours of the Higgs boson: A study in creativity and science motivation among high-school students in Italy. *Smart Learning Environments*, *8*, 1–23. https://doi.org/10.1186/s40561-021-00169-4

*Ardura, D., & Pérez-Bitrián, A. (2018). The effect of motivation on the choice of chemistry in secondary schools: adaptation and validation of the Science Motivation Questionnaire II to Spanish students. *Chemistry Education Research and Practice*, *19*(3), 905–918. https://doi.org/10.1039/C8RP00098K

*Austin, A. C., Hammond, N. B., Barrows, N., Gould, D. L., & Gould, I. R. (2018). Relating motivation and student outcomes in general organic chemistry. *Chemistry Education Research and Practice*, *19*(1), 331–341. https://doi.org/10.1039/C7RP00182G

Barnes, L. L., Harp, D., & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger state-trait anxiety inventory. *Educational and Psychological Measurement, 62*(4), 603–618. https://doi.org/10.1177/0013164402062004005

Batı, K., & Irmak, Ş. (2024). A reliability generalization study of the STEM-CIS scale: Exploring moderator effects. *Current Psychology, 43*, 35453–35470. https://doi.org/10.1007/s12144-024-07036-8

Beretvas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement, 63*(1), 75–95. https://doi.org/10.1177/0013164402239318

*Bernard, R. A. (2018). *The relationship of formal reasoning, motivation, and conceptual change: A quantitative study of introductory biology students across the United States* [Unpublished doctoral dissertation]. Montana State University.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335–340. https://doi.org/10.3102/10769986027004335

*Bónus, L., Antal, E., & Korom, E. (2024). Digital game-based inquiry learning to improve eighth graders' inquiry skills in biology. *Journal of Science Education and Technology*, *33*, 1–17. https://doi.org/10.1007/s10956-024-10096-x

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

*Breland, H., Clark, C. M., Shaked, S., & Paquette-Smith, M. (2023). The benefits of participating in a learning assistant program on the metacognitive awareness and motivation of learning assistants. *CBE—Life Sciences Education*, *22*(3), ar30. https://doi.org/10.1187/cbe.22-08-0156

*Chen, Y. C., Terada, T., & Aguirre-Mendez, C. (2023). Action research to explore argumentative writing projects supported with online simulation for developing conceptual knowledge and motivation to learn. *Reading and Writing*, *36*(9), 2275–2317. https://doi.org/10.1007/s11145-022-10363-6

*Christiansen, M. D. (2015). *Science as a direct measure of academic hope: The validation of the science hope scale* [Unpublished doctoral dissertation]. Hofstra University.

*Chumbley, S., Hainline, M. S., Wells, T., & Haynes, J. C. (2022). Students' motivation to learn science through undergraduate-level agricultural coursework. *Journal of Agricultural Education*, *63*(1), 182–199. https://doi.org/10.5032/jae.2022.01182

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101–129.

*Conner, L. C. (2021). *Impact of flipped learning on high school students' motivation, cognitive load, and perceptions of flipped approach for learning science* [Unpublished doctoral dissertation]. University of South Carolina.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*, 481–489.

*de Souza, R. T., da Silva, M. D. S. B., Barbato, D. M., de Guzzi, M. E. R., & Kasseboehmer, A. C. (2022). Motivation to learn chemistry: A thorough analysis of the CMQ-II within the Brazilian context. *Chemistry Education Research and Practice*, *23*(4), 799–810. https://doi.org/10.1039/D2RP00107A

DeVellis, R. F. (2003). *Scale development: Theory and application*. Sage.

*Dixon, K., & Wendt, J. L. (2021). Science motivation and achievement among minority urban high school students: An examination of the flipped classroom model. *Journal of Science Education and Technology*, *30*(5), 642–657. https://doi.org/10.1007/s10956-021-09909-0

*Dixon, K. L. (2017). *The effect of the flipped classroom on urban high school students' motivation and academic achievement in a high school science course* [Unpublished doctoral dissertation]. Liberty University.

*Dong, Z., Li, M., Minstrell, J., & Cui, Y. (2020). Psychometric properties of Science Motivation Questionnaire II-Chinese version in two waves of longitudinal data. *Psychology in the Schools*, *57*(8), 1240–1256. https://doi.org/10.1002/pits.22370

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629–634.

Eser, M. T., & Doğan, N. (2023). Life satisfaction scale: A meta-analytic reliability generalization study in Turkey sample. *Turkish Psychological Counseling and Guidance Journal, 13*(69), 224–239. https://doi.org/10.17066/tpdrd.1223320mn

*Formanek, M., Buxner, S., Impey, C., & Wenger, M. (2019). Relationship between learners' motivation and course engagement in an astronomy massive open online course. *Physical Review Physics Education Research*, *15*(2), 020140. https://doi.org/10.1103/PhysRevPhysEducRes.15.020140

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various coefficients of interrater reliability and agreement (Version 0.84.1) [R package]. Comprehensive R Archive Network. https://CRAN.R-project.org/package=irr

Glynn, S. M., Taasoobshirazi, G., & Brickman, P. (2009). Science motivation questionnaire: Construct validation with nonscience majors. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 46*(2), 127–146. https://doi.org/10.1002/tea.20267

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. https://doi.org/10.1002/tea.20442

*Gomes de Freitas, R. T., Tolentino Cicuto, C. A., & Pazinato, M. S. (2020). Motivation to learn natural sciences in the context of a training of teacher's course. *Problems of Education in the 21st Century*, *78*(3), 359–370. https://doi.org/10.33225/pec/20.78.359

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika, 41*(2), 219–231.

*Haviz, M., Maris, I. M., Azis, D., & Helmita, R. (2024). An investigation into science motivation, technology acceptance, and satisfaction intention during the transition to e-learning of prospective biology teachers. *Cogent Education*, *11*(1), 2393065. https://doi.org/10.1080/2331186X.2024.2393065

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504.

*Heim, A. B. (2020). *Understanding learner-centeredness and student engagement in undergraduate biology education* [Unpublished doctoral dissertation]. University of Northern Colorado.

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement and Evaluation in Counseling and Development, 35*(2), 113–127. https://doi.org/10.1080/07481756.2002.12069054

*Hibbard, L., Sung, S., & Wells, B. (2016). Examining the effectiveness of a semi-self-paced flipped learning format in a college general chemistry sequence. *Journal of Chemical Education*, *93*(1), 24–30. https://doi.org/10.1021/acs.jchemed.5b00592

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical Research Ed.), 327*(7414), 557–560.

Howell, R. T., & Shields, A. L. (2008). The file drawer problem in reliability generalization: A strategy to compute a fail-safe N with reliability coefficients. *Educational and Psychological Measurement, 68*(1), 120–128.

*Huda, N., & Rohaeti, E. (2023). The factors that influence the motivation to learn chemistry of upper-secondary school students in Indonesia. *Journal of Baltic Science Education*, *22*(4), 615–630. https://doi.org/10.33225/jbse/23.22.615

Huedo-Medina, T. B., Sánchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I² index? *Psychological Methods, 11*(2), 193–206. https://doi.org/10.1037/1082-989X.11.2.193

*Hugerat, M., Kortam, N., Maroun, N. T., & Basheer, A. (2020). The educational effectiveness of didactical games in project-based science learning among 5th grade students. *Eurasia Journal of Mathematics, Science and Technology Education*, *16*(10), em1888. https://doi.org/10.29333/ejmste/8490

Husain, W., Ebrahim, A. H., Alobaidi, S., Ghazzawi, H., Trabelsi, K., Pourdeh, E. F., ... & Jahrami, H. (2025). Reliability generalization meta-analysis of the internal consistency and test-retest reliability of the Body Awareness Questionnaire. *Journal of Health Psychology*. https://doi.org/10.1177/135910532513717

*Iyer, R. D. (2017). Uncovering students' motivation for learning science: Gender differences in Mumbai. *Global Business and Management Research*, *9*(3), 16.

Jian-Xin, Y., Yu-Xuan, X., Tian, L., Chu-Fan, D., Yu-Ying, G., & Fortus, D. (2023). Disciplinary learning motivation and its external influencing factors: Taking physics in a "selection crisis" as an example. *Research in Science Education, 53*(4), 823–839. https://doi.org/10.1007/s11165-023-10112-x

*Kahraman, S. (2023). The use of dynamic computer visualizations integrated with the poe sequence: its effect on learners' understanding, retention, and motivation. *Canadian Journal of Science, Mathematics and Technology Education*, *23*(2), 179–209. https://doi.org/10.1007/s42330-023-00284-z

*Kulesza, A. E., Imtiaz, S., & Bernot, K. M. (2022). Building connections to biology and community through service-learning and research experiences. *Journal of Microbiology & Biology Education*, *23*(3), e00082–22. https://doi.org/10.1128/jmbe.00082-22

*Kwarikunda, D., Schiefele, U., Ssenyonga, J., & Muwonge, C. M. (2020). The relationship between motivation for, and interest in, learning physics among lower secondary school students in Uganda. *African Journal of Research in Mathematics, Science and Technology Education*, *24*(3), 435–446. https://doi.org/10.1080/18117295.2020.1841961

*Kwarikunda, D., Schiefele, U., Ssenyonga, J., & Muwonge, C. M. (2021). Secondary school students' motivation profiles for physics learning: Relations with cognitive learning strategies, gender, attitudes and individual interest. *African Journal of Research in Mathematics, Science and Technology Education*, *25*(2), 197–210. https://doi.org/10.1080/18117295.2021.1956720

*Lawson, D. (2020). *Motivating critical thinkers in fourth grade science: Action research promoting 21st century skills through technology integrated project-based learning* [Unpublished doctoral dissertation]. University of South Carolina.

*Lee, G. G., & Mun, S. (2023). From science motivation to science identity: The mediating effect of science achievement according to gender. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(10), em2341. https://doi.org/10.29333/ejmste/13633

*Lei, Y. (2024). Unlocking the secrets of STEM success: Exploring the interplay of motivation to learn science, self-regulation, and emotional intelligence from a perspective of self-determination theory. *Learning and Motivation*, *87*, 102012. https://doi.org/10.1016/j.lmot.2024.102012

López-Ibáñez, C., López-Nicolás, R., Blázquez-Rincón, D. M., & Sánchez-Meca, J. (2024). Reliability generalization meta-analysis: Comparing different statistical methods. *Current Psychology, 43*(20), 18275–18293. https://doi.org/10.1007/s12144-023-05604-y

Mack, D. E., Vo, K. T., & Wilson, P. M. (2024). The long and short-form Warwick-Edinburgh mental wellbeing scale: a reliability generalization meta-analysis. *Journal of Happiness Studies, 25*(1), 12. https://doi.org/10.1007/s10902-024-00715-0

Meulenbroeks, R., van Rijn, R., & Reijerkerk, M. (2024). Fostering secondary school science students' intrinsic motivation by inquiry-based learning. *Research in Science Education, 54*(3), 339–358. https://doi.org/10.1007/s11165-023-10139-0

*Migalang, G. M. M., & Azuelo, A. G. (2020). Students' cognitive learning and motivation through hybrid instructional strategy. *Journal of Science & Mathematics Education in Southeast Asia*, *43*.

*Moro, K. C., & Billote, W. J. S. M. (2023). Integrating Ivatan indigenous games to learning module in physics: its effect to student understanding, motivation, attitude, and scientific sublime. *Science Education International*, *34*(1), 3–14. https://doi.org/10.33828/sei.v34.i1.1

Ng, D. T. K., & Chu, S. K. W. (2021). Motivating students to learn STEM via engaging flight simulation activities. *Journal of Science Education and Technology, 30*(5), 608–629. https://doi.org/10.1007/s10956-021-09907-2

Organization for Economic Cooperation and Development (OECD). (2007). *PISA 2006: Science competencies for tomorrow's world* (Volume I: Analysis). Author.

Park, S., & Kim, S. (2021). Is sustainable online learning possible with gamification? The effect of gamified online learning on student learning. *Sustainability, 13*(8), Article 4267. https://doi.org/10.3390/su13084267

*Pope, A. S. (2018). *Science attrition among high-performing college students: An analysis of motivational factors to improve retention strategies* [Unpublised doctoral dissertation]. Tennessee State University.

R Core Team. (2025). *A language and environment for statistical computing*. R foundation for statistical computing. https://www.R-project.org

*Rachmatullah, A., Roshayanti, F., Shin, S., Lee, J. K., & Ha, M. (2018). The secondary-student science learning motivation in Korea and Indonesia. *EURASIA Journal of Mathematics, Science and Technology Education*, *14*(7), 3123–3141. https://doi.org/10.29333/ejmste/91665

Rachmatullah, A., & Ha, M. (2019). Indonesian and Korean high school students' disparities in science learning orientations: An approach to multi-group structural equation modeling. *Asia-Pacific Science Education, 5*(1), 1–17. https://doi.org/10.1186/s41029-019-0048-5

Rahmayanti, F., Irmagita, A., Wardhany, I. I., & Gunawan, B. (2020). Measurement of oral medicine learning motivation in dental students using the Indonesian version of Science Motivation Questionnaire II (SMQ-II). *Pesquisa Brasileira Em Odontopediatria e Clínica Integrada, 20*, Article e5486. https://doi.org/10.1590/pboci.2020.136

*Razali, F., Manaf, U. K. A., & Ayub, A. F. M. (2020). STEM education in Malaysia towards developing a human capital through motivating science subject. *International Journal of Learning, Teaching and Educational Research*, *19*(5), 411–422. https://doi.org/10.26803/ijlter.19.5.25

*Razali, F. (2021). Exploring crucial factors of an interest in STEM career model among secondary school students. *International Journal of Instruction, 14*(2), 385–404. https://doi.org/10.29333/iji.2021.14222a

*Reece, A. J., & Butler, M. B. (2017). Virtually the same: A comparison of STEM students' content knowledge, course performance, and motivation to learn in virtual and face-to-face introductory biology laboratories. *Journal of College Science Teaching*, *46*(3), 83–89. https://doi.org/10.2505/4/jcst17_046_03_83

*Riccitelli, M. (2015). *Science identity's influence on Community College Students' engagement, persistence, and performance in biology* [Unpublished doctoral dissertation]. Northcentral University.

Roberts, D. (2007). Scientific literacy/science literacy. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 729–780). Lawrence.

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*, 306–322.

*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. https://doi.org/10.1039/C4RP00196F

*Salta, K., & Koulougliotis, D. (2020). Domain specificity of motivation: Chemistry and physics learning among undergraduate students of three academic majors. *International Journal of Science Education*, *42*(2), 253–270. https://doi.org/10.1080/09500693.2019.1708511

Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology, 66*(3), 402–425. https://doi.org/10.1111/j.2044-8317.2012.02057.doi:10.1111/j.2044-8317.2012.02057.x

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., & López-Nicolás, R. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods, 12*(4), 516–536. https://doi.org/10.1002/jrsm.1487

Sanchez-Meca, J., Rubio-Aparicio, M., López-Pina, J., Nuñez-Nuñez, R.M., & Marın-Martınez, F. (2015). *The phenomenon of reliability induction in the social and health sciences.* Paper presented at XIV Congress of Methodology of the Social and Health Sciences.

Sandoval-Lentisco, A., López-Nicolás, R., López-López, J. A., & Sánchez-Meca, J. (2023). Florida obsessive-compulsive inventory and children's Florida obsessive compulsive inventory: A reliability generalization meta-analysis. *Journal of Clinical Psychology, 79*(1), 28–42. https://doi.org/10.1002/jclp.23416

Scherer, R., & Teo, T. (2020). A tutorial on the meta-analytic structural equation modeling of reliability coefficients. *Psychological Methods, 25*(6), 747-775.

*Schmid, S., & Bogner, F. X. (2017). How an inquiry-based classroom lesson intervenes in science efficacy, career-orientation and self-determination. *International Journal of Science Education*, *39*(17), 2342–2360. https://doi.org/10.1080/09500693.2017.1380332

*Schönfelder, M. L., & Bogner, F. X. (2020). Between science education and environmental education: How science motivation relates to environmental values. *Sustainability*, *12*(5), 1968. https://doi.org/10.3390/su12051968

Schumm, M. F., & Bogner, F. X. (2016). Measuring adolescent science motivation. *International Journal of Science Education, 38*(3), 434–449. https://doi.org/10.1080/09500693.2016.1147659

Shin, D. D., Lee, M., Jung, S. J., & Bong, M. (2023). Relative effects of classroom utility value intervention on the science motivation of girls and boys. *Research in Science Education, 53*(3), 593–612. https://doi.org/10.1007/s11165-022-10070-w

*Shin, S., Rachmatullah, A., Ha, M., & Lee, J. K. (2018). A longitudinal trajectory of science learning motivation in Korean high school students. *Journal of Baltic Science Education*, *17*(4), 674–687. https://doi.org/10.33225/jbse/18.17.674

Thompson, B. (2002). *Score reliability: Contemporary thinking on reliability issues*. Sage.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*(2), 174–195. https://doi.org/10.1177/0013164400602

*Toma, R. B., Cordeiro Bizerra, A. M., Yánez, I., & Meneses Villagrá, J. Á. (2023). Cross-cultural adaptation of the Science Motivation Questionnaire II (SMQ-II) for Portuguese-speaking Brazilian secondary school students. *Revista Latinoamericana de Psicología*, *55*, 109–119. https://doi.org/10.14349/rlp.2023.v55.13

Tosun, C. (2013). Kimya Motivasyon Anketi-II'nin Türkçeye uyarlanması [Adaptation of the Chemistry Motivation Questionnaire-II into Turkish.]. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi, 15*(1), 173–202.

Tsui, H. K. H., Wong, T. Y., Ma, C. F., Wong, T. E., Hsiao, J., & Chan, S. K. W. (2025). Reliability of theory of mind tasks in schizophrenia, ASD, and nonclinical populations: A systematic review and reliability generalization meta-analysis. *Neuropsychology Review, 35*(3), 552–568. https://doi.org/10.1007/s11065-024-09652-4

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*(1), 6–20.

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development, 44*, 159–168. https://doi.org/10.1177/0748175611409845

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*(4), 509–522. https://doi.org/10.1177/00131640021970682

Van Vo, D., & Csapó, B. (2022). Exploring students' science motivation across grade levels and the role of inductive reasoning in science motivation. *European Journal of Psychology of Education, 37*(3), 807–829. https://doi.org/10.1007/s10212-021-00568-8

Viechtbauer, W. (2010). Conducting meta-analyses in R with the *metafor* package. *Journal of Statistical Software, 36*, 1–48.

*Watt, K. (2020). *Guided simulation: A new model of game-based pedagogy for non-stem students in the community college* [Unpublished doctoral dissertation]. Gwynedd Mercy University.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.

*Willoughby, B. R., Flack, N. A., Bird, R. J., & Woodley, S. J. (2024). Motivation to learn in university science students studying anatomy: A mixed-methods analysis of what drives learning. *Anatomical Sciences Education*, *17*(4), 818–830. https://doi.org/10.1002/ase.2416

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60*(2), 201–223.

*You, H. S., Kim, K., Black, K., & Min, K. W. (2018). Assessing science motivation for college students: Validation of the Science Motivation Questionnaire II using the rasch-andrich rating scale model. *Eurasia Journal of Mathematics, Science and Technology Education*, *14*(4), 1161–1173. https://doi.org/10.29333/ejmste/81821

*Zhdanov, S. P., Ishmuradova, A. M., Zakharova, V. L., Belous, S. V., & Grishnova, E. E. (2022). Validation and adaptation of the questionnaire on science motivation in the Russian context. *Eurasia Journal of Mathematics, Science and Technology Education*, *18*(8), em2136. https://doi.org/10.29333/ejmste/12219