Reliability generalization for the Motivated Strategies for Learning Questionnaire: A meta-analytic view of reliability estimates

David F. Holland
Amanda Kraha
Linda R. Zientek
Kim Nimon

University of Texas at Tyler, knimon@uttyler.edu
Julia A. Fulmore

See next page for additional authors

Follow this and additional works at: https://scholarworks.uttyler.edu/hrd_fac

Part of the Training and Development Commons

Recommended Citation
http://hdl.handle.net/10950/2324

This Article is brought to you for free and open access by the Human Resource Development at Scholar Works at UT Tyler. It has been accepted for inclusion in Human Resource Development Faculty Publications and Presentations by an authorized administrator of Scholar Works at UT Tyler. For more information, please contact tguillings@uttyler.edu.
Reliability Generalization of the Motivated Strategies for Learning Questionnaire: A Meta-Analytic View of Reliability Estimates

David F. Holland¹, Amanda Kraha², Linda R. Zientek³, Kim Nimon⁴, Julia A. Fulmore⁵, Ursula Y. Johnson⁶, Hector F. Ponce⁷, Mariya Gavrílova Aguilar⁸, and Robin K. Henson⁸

Abstract
A reliability generalization meta-analysis was performed to explore the relationship between study factors and levels of alpha reliability for the 15 subscales of the Motivated Strategies for Learning Questionnaire (MSLQ). The MSLQ has been widely adapted over the past 25 years to investigate the role of motivation and strategies in learning, primarily at the postsecondary level. A literature search from the years 1991 to 2015 yielded 295 peer-reviewed journal articles and 1,369 alpha reliability coefficients. Articles were coded for six potential moderator study variables. A novel varying coefficient (VC) model was adopted to determine average reliabilities across studies for each subscale and to perform multiple regression analyses to identify study variables that may moderate alpha reliability estimates. Commonality analyses were used to aid in interpretation of regression results. Meta-analyzed alpha reliabilities were lower than values published in the test manual for all but three of the subscales. Ability of specific moderators to predict score reliability varied across subscales; however, studies in North America, in English, or using a 7-point response scale generally corresponded to increased reliability estimates. Knowledge of expected levels of score reliability under varying sample and study conditions may provide useful information for researchers planning future use of the MSLQ.

Keywords
educational research, education, social sciences, reliability and validity, research methods, social sciences, Motivated Strategies for Learning Questionnaire, measurement and scaling methods, research methodology and design

Student motivation and learning strategies have long been recognized as important predictors to student success. Prior to the mid-1980s, research in this area focused on individual differences and learning styles without providing clear connections to the manner in which students acquire new knowledge through behaviors and cognition (Duncan & McKeachie, 2005). Over time, a recognition emerged that learning occurs through an interplay of various internal and external sources. In response, Pintrich, Smith, Garcia, and McKeachie (1991, 1993) developed the Motivated Strategies for Learning Questionnaire (MSLQ), which was based on a self-regulated learning perspective, taking into account cognitive, motivational, and behavioral factors of learning within a social context (Pintrich, 2004). Thus, the development of the MSLQ grew out of a pragmatic need for a theoretically based instrument for use by faculty and students alike to improve post-secondary learning. Subsequently, the MSLQ has served to emphasize the interplay between cognition and motivation in learning, helped to operationalize the constructs of self-regulated learning theory, and provided a useful tool to perform empirical investigations of these constructs (Credé & Phillips, 2011; Duncan & McKeachie, 2005).
Prevalence of Use

The MSLQ has enjoyed widespread use in the evaluation of the effects of course designs on student learning, characterization of motivation and use of learning strategies across various target populations, and exploration of motivational constructs and individual differences in self-regulated learning (Duncan & McKeachie, 2005). The instrument has been widely utilized in the United States and internationally, having been translated into at least 11 languages other than English. Although the MSLQ was designed for use in a postsecondary environment, subscales of the instrument have been used in settings ranging from elementary school (e.g., Andreou & Metallidou, 2004) through adult online education (e.g., Richardson, 2007). Use of the MSLQ has varied widely across research fields, study designs, and populations and on the subscale components administered. The popularity of administering the MSLQ might be attributed to the fact that it is in the public domain, and that the modular nature of the instrument allows for easy use of some or all of the 15 subscales, depending on the needs of the researcher. Indeed, based on Duncan and McKeachie’s (2005) sample of 56 empirical studies using the MSLQ, researchers appear more frequently to use only portions of the instrument, rather than the entire scale.

Structure of the MSLQ

The college version of the MSLQ (Pintrich et al., 1991, 1993) is an 81-item self-report instrument containing 15 subscales divided into motivation and learning strategies sections. Items are scored on a Likert-type scale anchored at 1 (not at all true of me) and 7 (very true of me), and scale scores are based on means across items in the scale, rather than sum totals. After several modifications, the final version was administered in 1990 to 380 college students in the Midwestern United States and scores were subjected to psychometric analysis (Pintrich et al., 1991).

The organizational structure of the MSLQ is provided in Table 1. The motivational scales consist of 31 items addressing three theoretical components of motivation: value beliefs, expectancy, and affect (Duncan & McKeachie, 2005). Value beliefs are assessed with three subscales pertaining to Intrinsic Goal Orientation (four items), Extrinsic Goal Orientation (four items), and Task Value (six items). Expectancy is assessed with two subscales, Self-Efficacy (eight items) and Control Beliefs About Learning (four items). Finally, affect is assessed through a single subscale concerning Test Anxiety (five items). As reported in the test manual (Pintrich et al., 1991), alpha reliabilities for scores obtained from the 380-student sample ranged from .62 for the Extrinsic Goal Orientation subscale to .93 for the Self-Efficacy subscale (Table 1).
The learning strategies scales consist of 50 items, which include nine subscales addressing cognitive, meta-cognitive, and resource management strategies. Cognitive strategies are assessed with four subscales measuring Rehearsal (four items), Elaboration (six items), Organization (four items), and Critical Thinking (five items). Meta-cognitive strategies are evaluated with a single, 12-item subscale. The final four subscales address aspects of resource management and include Time and Study Environment Regulation (eight items), Effort Regulation (four items), Peer Learning (three items), and Help Seeking (four items). Alpha reliability estimates for the learning strategies subscales provided in the MSLQ test manual are generally lower than those for the motivational scales, ranging from .52 for the Help Seeking subscale to .80 for the Critical Thinking subscale (Table 1).

Reliability and the MSLQ

Despite the widespread use of the MSLQ, several concerns have been raised about the psychometric properties of the instrument. In terms of the internal consistency reliability estimates obtained for the various subscales, Pintrich et al. (1993) claimed that the “coefficient alphas for the motivational scales are robust, demonstrating good internal consistency” (p. 808) and “the alphas for the learning strategies scales are reasonable” (p. 809). However, they did not explain by what evaluative standard they are determined to be “good” or “reasonable.” Although the authors may have made their robustness determination based on comparison data collected during development of the MSLQ, it is well known that reliability estimates such as coefficient alpha vary with changing sample characteristics, study conditions, and score distributions (Crocker & Algina, 1986; Pedhazur & Schmelkin, 1991; Thompson, 2003; Wilkinson & American Psychological Association [APA] Task Force on Statistical Inference, 1999). Nunnally (1978) provided some often-cited rules of thumb to determine the adequacy of levels of reliability suggesting that “reliabilities of .70 or higher will suffice” when “in early stages of research on predictor tests or hypothesized measures or a construct” (p. 245) and that reliabilities of at least .80 are appropriate for basic research purposes. However, of the 15 MSLQ subscales for which Pintrich et al. (1993) reported reliabilities, nine subscales (60%) demonstrated sufficient score reliability estimates to meet the .70 standard for introductory research and three subscales fell at the .68 or .69 level. Only four subscales (27%) met the more appropriate standard of .80 for basic research in theory testing with an additional subscale at the .79 level.

Due to its modular nature and ease of administration, researchers have routinely utilized instrument sections (e.g., Nielsen, 2004), scales (e.g., Arend, 2007), subscales (e.g., Hodges & Kim, 2010), and individual test items (e.g., Husman & Hilpert, 2007) of the MSLQ to fit their particular research needs. However, caution should be exercised when selecting items in this manner as researchers should not assume that psychometric properties of instrument components remain consistent across various applications, study designs, samples, and time. For example, MSLQ subscale alpha reliability estimates obtained by Hilpert, Stempien, van der Hoeven Kraft, and Husman (2013) consistently met or exceeded those reported by Pintrich et al. (1991) in the MSLQ manual (Table 1). For three subscales, Extrinsic Goal Orientation, Control of Learning Beliefs, and Help Seeking, the differences were .10 or higher, although both sets of estimates were based on data collected from similar samples of undergraduate college students in the United States. For each study, the consistency of scores as measured by reliability coefficients will vary for different sample characteristics. Thus, a study is warranted that meta-analytically examines the predictors of reliability coefficients for the MSLQ when administered across a variety of samples.

Reliability Generalization (RG)

In general, score validity concerns the degree of trustworthiness of inferences made from the data collected, and to evaluate score validity the consistency of measurement, or reliability, must also be known. Thus, when performing substantive studies, researchers seek to utilize instruments that consistently and accurately measure constructs of interest and a failure to do so may lead to false conclusions (Thompson, 2003).

It has long been known that estimates of reliability vary with changing sample characteristics, study conditions, and score distributions (Crocker & Algina, 1986; Pedhazur & Schmelkin, 1991; Thompson, 2003). For this reason, journal editors (Thompson, 1994; Vacha-Haase, Henson, & Caruso, 2002) and professional organizations (American Educational Research Association, APA, & National Council on Measurement in Education, 1999; APA, 2001; Wilkinson & APA Task Force on Statistical Inference, 1999) have advocated that authors always provide reliability estimates for the data in hand. Such transparent reporting practice encourages researchers to take score reliability into account when interpreting study results, and provides consumers of research critical information necessary to make informed judgments regarding the viability of data interpretations and study conclusions. In instances where primary researchers appropriately have reported psychometric data for administration of a particular instrument, it may be desirable to examine score reliability on multiple occasions to discern how measurement error may vary under fluctuating study conditions. Such an approach requires a quantitative integration of reliability coefficients, which is best suited for meta-analytic methods such as RG (see Sánchez-Meca, López-López, & López-Pina, 2013; Vacha-Haase, 1998; Vacha-Haase & Thompson, 2011).

RG is a meta-analytic method for synthesizing reliability coefficients across studies (Caruso, 2000; Vacha-Haase, 2001).
and is “used to explore variability in reliability estimates and characterize the sources of this variance” (Vacha-Haase et al., 2002, p. 562). RG studies provide insight into the nature of score reliability in prior applications of a test, which may help future researchers estimate expected levels of measurement error and inform study design decisions regarding effect sizes, power, and statistical significance (Henson & Thompson, 2002; Nimon, Zientek, & Henson, 2012). It is recommended that authors cite available RG results when describing tests used in substantive studies to provide comparative data to facilitate interpretation of outcomes (Leech, Onwuegbuzie, & O’Conner, 2011). Bonett (2010) encourages researchers planning use of a test to perform a preliminary RG on a small number of carefully selected, high-quality studies to obtain more accurate estimates of expected reliability and to identify potential effects of moderator variables. Such retrospective and prospective practices promote meta-analytic thinking, which serves to build a historical contextual framework in which to better evaluate single-study outcomes (Cumming & Finch, 2001; Henson, 2006; Thompson, 2002). As Bonett (2010) notes, “the use of meta-analysis to statistically incorporate prior information into a current study has the potential to revolutionize behavioral research and help achieve the goals of an integrative and cumulative science” (p. 380).

**RG Method**

Since 1998, well more than 100 RG studies have been published on an assortment of psychological instruments employing a wide variety of meta-analytic and statistical methods; however, there appears to be no firmly established best practice when performing such studies (Holland, 2015; Sánchez-Meca et al., 2013). Several methodological decisions must be made by the RG meta-analyst, including selection of statistical models for coefficient synthesis and moderator analysis, and the transformation and weighting of coefficients within these models.

Two classes of statistical models traditionally have been used in meta-analyses: the fixed-effects (FE) model of Hedges and Olkin (1985) and the random-effects (RE) models of Hedges and Vevea (1998) or Hunter and Schmidt (2004). Classical FE models are based on the assumption that study coefficients are all estimating the same population parameter, and any deviation from the parameter is the result of sampling error (Bonett, 2010; Hedges, 1992). In general, FE models are recommended when one wishes to generalize the results to studies similar to those included in the meta-analysis. FE methods have been determined to exhibit poor performance under conditions typical of many meta-analyses and are generally not recommended for routine use (Bonett, 2008; Hunter & Schmidt, 2004; Rodriguez & Maeda, 2006; Schmidt, Oh, & Hayes, 2009).

RE statistical models are based on the assumption that multiple population parameters exist, and that each study included in the meta-analysis represents a sample of a hypothetical population of past or future studies. Thus, each coefficient is considered to be an estimate of its own population parameter, which may vary from study to study. RE models include two error components in synthesized reliability estimates: the within-study variance and the between-study variance. Due to the additional error accounted for by the between-study variance component, RE models tend to produce wider confidence intervals (CIs) than FE models when synthesizing reliability coefficients across studies (Sánchez-Meca et al., 2013). The application of an RE model in the meta-analysis of coefficient alpha by Rodriguez and Maeda (2006) has been called into question based on the introduction of bias in parameter estimates, lack of interpretable estimates of parameter variance, and violations of sampling assumptions of the model (Bonett, 2010).

First proposed for use in meta-analysis by Laird and Mosteller (1990), a VC statistical model has been applied by Bonett (2010) to the meta-analysis of coefficient alpha. The VC model provides an alternative to traditional FE or RE models, retaining beneficial characteristics of both approaches. As a type of FE model, results from the VC analysis may be generalized only to studies similar to those included in the meta-analysis. However, rather than assuming that alpha estimates are all equal to a single fixed parameter, each study is assumed to estimate its own population reliability coefficient, similar to the RE approach. The magnitude of error components are moderate under the VC model and produce CIs intermediate between those estimated under FE or RE models (Sánchez-Meca et al., 2013). The VC model has excellent small-sample performance characteristics in parameter estimation, provides more accurate CIs, and can be used over a much wider range of problems than traditional models (Bonett, 2010). For these reasons, Bonett’s VC model was utilized for the current study.

To synthesize coefficients across studies, Bonett (2010) recommends calculation of the simple arithmetic mean of unweighted, untransformed alphas. To derive CIs for these means, the VC model utilizes a log-complement transformation, \(\ln(1 - \alpha_j)\), where \(\alpha_j\) is the alpha estimate of study \(j\), to stabilize variance and normalize the distribution of alpha, and applies the delta method to estimate variance from each study. Individual study variances, are then used to determine CIs for the mean (see Bonett, 2010). Krizan (2010) has developed an Excel worksheet for calculation of means and CIs based on Bonett’s proposed methods.

Bonett (2010) recommends use of ordinary least squares (OLS) linear regression to investigate the potential effects of both categorical and continuous moderator variables, using transformed alpha, \(\ln(1 - \alpha_j)\), as the outcome variable. The linear function may be expressed as

\[
p = Xb + \varepsilon,
\]
where \( p \) is a vector of log-complement transformed alphas, expressed as \( \ln(1 - \alpha_j) - \ln(n / [n - 1]) \), where the second term is a correction factor. \( X \) represents a design matrix of \( k \) potential study moderator variables, \( b \) represents a vector of unknown parameters, and \( \epsilon \) is a vector of random sampling errors, such that \( \text{var}(\epsilon_j) = \text{var}(\ln(1 - \alpha_j)) \). OLS estimates for \( b \) coefficients are determined as

\[
b = (XX)^{-1} X'p
\]

and the covariance matrix estimated as

\[
cov(b) = (XX)^{-1} VXVX(XX)^{-1},
\]

where \( V \) is a diagonal matrix with \( \text{var}(\epsilon) \) as the \( j \)th element (see Bonett, 2010, for more details). SPSS syntax to perform this moderator analysis is provided in Appendix A.

Because the dependent variable in this model is based on a normalizing transformation of alpha, it is suggested that \( b_j \) be back-transformed as \( \exp(b) \) to improve interpretation of regression coefficients. In this manner, \( \exp(b) \) may be interpreted as “the multiplicative change in nonreliability for every 1-point increase in the \( k \)th predictor variable while the values of all other predictor variables are held constant” (Bonett, 2010, p. 372). Thus, \( \exp(b) \) values less than 1.0 indicate that the predictor variable is related to decreases in unreliability (i.e., increases in reliability), taking into account all other predictors. Development of a regression model relating study moderator variables to reliability estimates may allow researchers to predict expected values of coefficient alpha in future studies, given known values of predictor variables.

**Purpose**

The purpose of the current study is to perform an RG meta-analysis to explore the variability of MSLQ subscale score reliability across studies and to determine the potential relationship between study factors and the variability of subscale reliability. In light of the contributions and widespread use of the MSLQ in research surrounding learning motivation theory, an evaluation of factors that predict measurement reliability from MSLQ administrations would be beneficial for researchers who are contemplating using the MSLQ in the future. Results from an RG study will be of value to researchers who will be able to make educated and informed decisions when planning their study on motivational and learning strategies for their given sample.

**Method**

Peer-reviewed journal articles utilizing the MSLQ were collected in two waves. In September 2010, articles were identified using the online search engine, Google Scholar, with filters set to return works published in the years 1991 through 2010 in which either of the two seminal MSLQ publications (i.e., Pintrich et al., 1991, 1993) were cited. In an effort to capture all citations from 2010, this process was repeated in July 2013. The two searches returned 903 unique citations for which articles were procured. Of the available sources, 315 studies were published in peer-reviewed journals and administered one or more subscales of the MSLQ. These articles were inspected for reporting of reliability coefficients. A total of 168 studies were removed from the analysis due to insufficient reliability reporting, including failure to report reliability for data collected, or reporting coefficients in an unusable format, such as ranges of values over several subscales. Alpha coefficients from the remaining 147 articles were collected and study variables coded. The second wave of data collection was performed in September 2017, whereby journal articles published from 2011 to 2015 were identified utilizing Publish or Perish (Version 5; Harzing, 2016), a software program that retrieves citations from Google Scholar. Separate queries using the Lookup Citations function were conducted for each of the two Pintrich et al. (1991, 1993) seminal articles. A total of 625 citations were retrieved and after the removal of duplicates and books, 545 citations remained. Sources for citations were obtained and, on further inspection, 276 studies were not available, not peer-reviewed journal articles, could not be translated, or did not utilize the MSLQ. Of the remaining 269 articles utilizing the MSLQ, 121 did not report alpha reliability coefficients in a manner suitable for meta-analysis, leaving 148 articles suitable for the current study.

Combining articles from both waves of data collection, reliability coefficients from a total of 295 articles were available for further analysis. Thus, among the 584 studies reporting use of the MSLQ, only 51% provided alpha reliability coefficients for the data in hand. We acknowledge that focusing on peer-reviewed journals may potentially create a publication bias—however, we believe it is the most efficient way to focus our search for this popularly used measure. A list of articles included in the study is provided in Appendix B.

**Coding**

Articles initially were coded by four trained raters and then two additional raters were added for the second wave. Coding was verified by two of the authors, who reached agreement in cases where opinions differed. Multiple reliability estimates from a study were utilized if they were obtained from distinct samples. In cases where several estimates were reported for the same sample over multiple administrations (e.g., pretest/posttest), only the first estimate was included in the analysis in an effort to maintain independence of observations (see Romano & Kromrey, 2009, for a discussion of independence issues in RG studies).

Components of the MSLQ have been applied internationally to a variety of research settings and applications that
may vary widely from the original studies performed by Pintrich and colleagues (1991) in the United States. Researchers have also freely modified MSLQ subscales by changing item wording to fit a particular need, adding or deleting items from scales, and translating the instrument into languages other than English. It is of interest in the current study to determine how such varying study applications and instrument modifications may relate to variability in score reliability.

An initial set of coding variables was selected based on those utilized in prior RG studies (Henson & Thompson, 2002; Vacha-Haase & Thompson, 2011), including participant, study, and instrument characteristics. From these, we selected key potential moderator variables that reflected typical modifications of the instrument, and from a practical standpoint, other features most likely to be reported in published studies as recommended by the American Educational Research Association (2006) and the APA (2010). Coded variables included the age, gender, and race/ethnicity of participants; the educational setting and location (country) of the study; and instrument characteristics, including number of response scale choices, wording modifications, and language of translation.

Following the coding process, the data set was inspected to determine which of the coded study variables were reported in sufficient numbers to adequately represent the sample of studies and support the planned moderator analyses. Of the 344 samples described in the primary studies, only 27% included information on participant race/ethnicity and only 62% included the mean age; hence, these predictors were excluded from the analysis. Categorical variables consisting of multiple levels, such as study location, educational setting, and instrument language, were collapsed into dichotomous variables to reduce the number of predictor variables in the model. Ultimately, five categorical variables were dichotomously coded to indicate whether or not the application was similar to the original study by Pintrich et al. (1991). These variables included use of a 7-point response scale, use of original item wording, use of an English version of the instrument, selecting a study population consisting of post-secondary students, and performing the study in North America. In addition a single quantitative variable, percentage of males in the sample was also coded and included in the moderator analysis.

**Results**

**Study Characteristics**

Alpha coefficients were reported for 344 unique samples across the 295 articles subjected to review. Although the majority of samples (67%) included undergraduate students, 32% included students from Grades 3 through 12, and 15% included students at a graduate level (Table 2). Studies included in the analysis were performed in 32 different countries, with 52% of samples originating in the United States and Canada; however, all continents were represented. In addition, study authors reported use of translations of the MSLQ into 14 different languages, most commonly Turkish, Dutch, Spanish, and Chinese. In 61% of the studies, researchers utilized the original 7-point Likert-type scale (Table 2).

Because researchers may have used one, several, or all the 15 MSLQ subscales in their studies, the number of reliability estimates obtained differed markedly across the subscales (Table 3). A total of 1,369 coefficients were meta-analyzed across all studies and subscales. Mean number of coefficients collected for each subscale was 91.2 ($SD = 38.6$). The most commonly represented subscales were Self-Efficacy, with 199 coefficients collected, and Metacognitive Self-Regulation, with 149 coefficients, whereas the fewest number of coefficients were collected for the Help Seeking and Peer Learning subscales, with 57 and 46 coefficients meta-analyzed, respectively.

**Mean Reliability Scores**

Following Bonett’s (2010) method, we derived an unweighted average estimator of alpha reliability across studies and a 95% CI for each subscale, based on all study coefficients available. Estimates ranged from .608 for Help Seeking to .879 for the Self-Efficacy subscale (Table 3). Synthesized alpha reliabilities were generally lower than those published in the MSLQ manual (Pintrich et al., 1991), with the exception of the Extrinsic Goal Orientation, Organization, and Help Seeking subscales. Two of the motivation subscales and five of the learning strategy subscales produced mean reliability scores of less than .70. CIs estimated with the VC approach were relatively narrow, with widths ranging from .005 for Self-Efficacy to .032 for Peer Learning. It was not surprising that the subscale with the greatest reliability estimate (Self-Efficacy) had the narrowest CI, as its estimation is dependent on the variance of the estimate, which tends to decrease with increasing reliability, the magnitude of reliability, as well as sample size, which was the largest of any of the subscales.

**Moderator Analysis**

A general linear model was utilized to examine moderator effects of sample and study characteristics on estimates of reliability. Using Bonett’s (2010) OLS multiple regression method, categorical and quantitative study variables were used as predictors of log-complement, bias-adjusted study reliability estimates.

Multiple regression methods require that all predictor and outcome variables are represented for all records in the data set. An issue commonly encountered in meta-analytic studies is that authors of primary research may not be fully transparent in their description of sample characteristics and study design; thus, data collection is often plagued by missing data.
for potential moderator variables (Thompson & Vacha-Haase, 2000). In this study, missing data were handled through listwise deletion, which had the potential to cause considerable reduction in sample size for some subscales. A summary of the coded predictor variables and number of missing values for each subscale is provided in Table 4. The predictor variable most commonly missing from the data set was percent male, with missing values ranging from 5.3% for the Critical Thinking scale to 21.3% for Test Anxiety, followed by the 7-point scale predictor, which had a maximum of 6.5% missing data for Peer Learning. Despite the loss of records due to missing data, final sample sizes were near or above 60 for all but the Peer Learning and Help Seeking scales. Final sample sizes and results of the OLS moderator analyses are provided in Table 5.

Commonality analysis aids in interpreting regression results by partitioning the total $R^2$ effect size into common and unique variance accounted for by the predictor variables (Nimon, Lewis, Kane, & Haynes, 2008; Zientek & Thompson, 2006). Commonality analyses allow for interpretation of a predictor variable’s contribution to the model both alone and in combination with other predictors. Results of commonality analysis for potential moderator variables are provided in Table 6.

OLS regression on transformed alpha coefficients revealed that reliabilities of all 15 subscales appeared to be moderated by at least one of the study variables, with the exception of Extrinsic Goal Orientation, and most subscales appeared to have several moderating variables. Postsecondary, 7-point response scale, and English most often influence

---

**Table 2. Study and Sample Characteristics.**

<table>
<thead>
<tr>
<th>Country</th>
<th>N studies</th>
<th>Language</th>
<th>N studies</th>
<th>Educational setting</th>
<th>N samples</th>
<th>Scale type</th>
<th>N studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>143</td>
<td>English</td>
<td>185</td>
<td>Undergraduate</td>
<td>229</td>
<td>7-point</td>
<td>181</td>
</tr>
<tr>
<td>Canada</td>
<td>36</td>
<td>Turkish</td>
<td>26</td>
<td>High school</td>
<td>49</td>
<td>5-point</td>
<td>77</td>
</tr>
<tr>
<td>Turkey</td>
<td>29</td>
<td>Dutch</td>
<td>12</td>
<td>Graduate</td>
<td>46</td>
<td>6-point</td>
<td>13</td>
</tr>
<tr>
<td>Australia/New Zealand</td>
<td>14</td>
<td>Spanish</td>
<td>12</td>
<td>Middle school</td>
<td>14</td>
<td>4-point</td>
<td>6</td>
</tr>
<tr>
<td>Singapore</td>
<td>11</td>
<td>Chinese</td>
<td>11</td>
<td>Nonstudent</td>
<td>7</td>
<td>100-point</td>
<td>1</td>
</tr>
<tr>
<td>Taiwan</td>
<td>11</td>
<td>German</td>
<td>7</td>
<td>Elementary</td>
<td>4</td>
<td>Not stated</td>
<td>17</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>11</td>
<td>Norwegian</td>
<td>7</td>
<td>Postgraduate</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td>10</td>
<td>French</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Netherlands</td>
<td>9</td>
<td>Arabic</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>9</td>
<td>Hebrew</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norway</td>
<td>8</td>
<td>Malay</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>7</td>
<td>Farsi</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td>5</td>
<td>Greek</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greece</td>
<td>5</td>
<td>Japanese</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korea</td>
<td>5</td>
<td>Slovenian</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Israel</td>
<td>4</td>
<td>Not stated</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oman</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>France</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malaysia</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argentina</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bahrain</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fiji</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philippines</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Columbia</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Croatia</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finland</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iran</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japan</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kuwait</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peru</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovenia</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>United Arab Republic</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not determined</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>346</td>
<td>300</td>
<td>353</td>
<td>295</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** The totals provided exceed the number of studies included in the meta-analysis due to studies conducted in multiple countries, using multiple languages, and samples derived from multiple educational settings.
Table 3. Summary Statistics for the Meta-Analysis of Alpha Reliability for the Subscales of the MSLQ.

<table>
<thead>
<tr>
<th>Subscale</th>
<th>m coeff</th>
<th>Total N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>M</th>
<th>LL</th>
<th>UL</th>
<th>Manual</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Motivation section</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intrinsic Goal Orientation</td>
<td>112</td>
<td>31,712</td>
<td>.37</td>
<td>.88</td>
<td>.709</td>
<td>.703</td>
<td>.719</td>
<td>.74</td>
</tr>
<tr>
<td>Extrinsic Goal Orientation</td>
<td>82</td>
<td>24,224</td>
<td>.48</td>
<td>.92</td>
<td>.692</td>
<td>.685</td>
<td>.703</td>
<td>.62</td>
</tr>
<tr>
<td>Task Value</td>
<td>105</td>
<td>36,276</td>
<td>.51</td>
<td>.95</td>
<td>.833</td>
<td>.829</td>
<td>.839</td>
<td>.90</td>
</tr>
<tr>
<td>Control Beliefs</td>
<td>75</td>
<td>21,888</td>
<td>.35</td>
<td>.87</td>
<td>.645</td>
<td>.635</td>
<td>.661</td>
<td>.68</td>
</tr>
<tr>
<td>Self-Efficacy for Learning</td>
<td>199</td>
<td>69,177</td>
<td>.48</td>
<td>.96</td>
<td>.879</td>
<td>.878</td>
<td>.883</td>
<td>.93</td>
</tr>
<tr>
<td>Test Anxiety</td>
<td>80</td>
<td>27,884</td>
<td>.56</td>
<td>.91</td>
<td>.759</td>
<td>.754</td>
<td>.767</td>
<td>.80</td>
</tr>
<tr>
<td><strong>Learning strategies section</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rehearsal</td>
<td>75</td>
<td>30,089</td>
<td>.24</td>
<td>.83</td>
<td>.668</td>
<td>.660</td>
<td>.680</td>
<td>.69</td>
</tr>
<tr>
<td>Elaboration</td>
<td>91</td>
<td>35,517</td>
<td>.41</td>
<td>.87</td>
<td>.745</td>
<td>.739</td>
<td>.753</td>
<td>.76</td>
</tr>
<tr>
<td>Organization</td>
<td>69</td>
<td>30,293</td>
<td>.24</td>
<td>.84</td>
<td>.679</td>
<td>.671</td>
<td>.691</td>
<td>.64</td>
</tr>
<tr>
<td>Critical Thinking</td>
<td>76</td>
<td>27,619</td>
<td>.44</td>
<td>.90</td>
<td>.778</td>
<td>.773</td>
<td>.786</td>
<td>.80</td>
</tr>
<tr>
<td>Metacognitive Self-Regulation</td>
<td>149</td>
<td>55,175</td>
<td>.50</td>
<td>.96</td>
<td>.754</td>
<td>.750</td>
<td>.762</td>
<td>.79</td>
</tr>
<tr>
<td>Time and Study Environment</td>
<td>68</td>
<td>23,183</td>
<td>.50</td>
<td>.85</td>
<td>.724</td>
<td>.718</td>
<td>.734</td>
<td>.76</td>
</tr>
<tr>
<td>Effort Regulation</td>
<td>85</td>
<td>28,495</td>
<td>.32</td>
<td>.85</td>
<td>.660</td>
<td>.652</td>
<td>.674</td>
<td>.69</td>
</tr>
<tr>
<td>Peer Learning</td>
<td>46</td>
<td>15,154</td>
<td>.41</td>
<td>.78</td>
<td>.628</td>
<td>.614</td>
<td>.646</td>
<td>.76</td>
</tr>
<tr>
<td>Help Seeking</td>
<td>57</td>
<td>16,804</td>
<td>.35</td>
<td>.86</td>
<td>.608</td>
<td>.596</td>
<td>.625</td>
<td>.52</td>
</tr>
</tbody>
</table>

Note. m coeff = number of alpha coefficients synthesized; total N = overall sample size across studies; M = unweighted mean estimate of alpha coefficients; LL and UL = lower and upper limits of 95% confidence interval; Manual = alpha coefficient estimates published in the MSLQ manual (Pintrich, Smith, Garcia, & McKeachie, 1991). MSLQ = Motivated Strategies for Learning Questionnaire.

Table 4. Summary of Predictor (Moderator) Variable Coding and Missing Data.

<table>
<thead>
<tr>
<th>Scale</th>
<th>m coeff</th>
<th>Coding</th>
<th>7-point scale</th>
<th>Original wording</th>
<th>English</th>
<th>Postsecondary</th>
<th>North America</th>
<th>Percent male</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intrinsic</strong></td>
<td>112</td>
<td>% yes</td>
<td>77.7</td>
<td>49.1</td>
<td>70.5</td>
<td>81.3</td>
<td>58.0</td>
<td>17.0</td>
</tr>
<tr>
<td><strong>Extrinsic</strong></td>
<td>82</td>
<td>% yes</td>
<td>76.8</td>
<td>48.8</td>
<td>64.6</td>
<td>85.4</td>
<td>47.6</td>
<td>20.7</td>
</tr>
<tr>
<td><strong>Task Value</strong></td>
<td>105</td>
<td>% yes</td>
<td>70.5</td>
<td>41.0</td>
<td>63.8</td>
<td>78.1</td>
<td>45.7</td>
<td>15.2</td>
</tr>
<tr>
<td><strong>Control of Learning Beliefs</strong></td>
<td>75</td>
<td>% yes</td>
<td>76.0</td>
<td>44.0</td>
<td>62.7</td>
<td>81.3</td>
<td>45.3</td>
<td>20.0</td>
</tr>
<tr>
<td><strong>Self-Efficacy</strong></td>
<td>199</td>
<td>% yes</td>
<td>66.3</td>
<td>36.7</td>
<td>64.3</td>
<td>75.4</td>
<td>49.2</td>
<td>12.1</td>
</tr>
<tr>
<td><strong>Test Anxiety</strong></td>
<td>80</td>
<td>% yes</td>
<td>70.0</td>
<td>43.8</td>
<td>62.5</td>
<td>75.0</td>
<td>40.0</td>
<td>21.3</td>
</tr>
<tr>
<td><strong>Rehearsal</strong></td>
<td>75</td>
<td>% yes</td>
<td>70.7</td>
<td>40.0</td>
<td>57.3</td>
<td>80.0</td>
<td>49.3</td>
<td>9.3</td>
</tr>
<tr>
<td><strong>Elaboration</strong></td>
<td>91</td>
<td>% yes</td>
<td>70.3</td>
<td>37.4</td>
<td>60.4</td>
<td>83.5</td>
<td>53.8</td>
<td>7.7</td>
</tr>
<tr>
<td><strong>Organization</strong></td>
<td>69</td>
<td>% yes</td>
<td>68.1</td>
<td>33.3</td>
<td>55.1</td>
<td>76.8</td>
<td>44.9</td>
<td>10.1</td>
</tr>
<tr>
<td><strong>Critical Thinking</strong></td>
<td>76</td>
<td>% yes</td>
<td>72.4</td>
<td>46.1</td>
<td>60.5</td>
<td>85.5</td>
<td>52.6</td>
<td>6.0</td>
</tr>
<tr>
<td><strong>Metacognition</strong></td>
<td>149</td>
<td>% yes</td>
<td>67.1</td>
<td>34.9</td>
<td>61.7</td>
<td>75.2</td>
<td>52.3</td>
<td>5.3</td>
</tr>
<tr>
<td><strong>Time Management</strong></td>
<td>68</td>
<td>% yes</td>
<td>67.6</td>
<td>36.8</td>
<td>52.9</td>
<td>85.3</td>
<td>52.3</td>
<td>4.5</td>
</tr>
<tr>
<td><strong>Effort Regulation</strong></td>
<td>85</td>
<td>% yes</td>
<td>67.1</td>
<td>37.6</td>
<td>60.0</td>
<td>70.6</td>
<td>51.8</td>
<td>8.8</td>
</tr>
<tr>
<td><strong>Peer Learning</strong></td>
<td>46</td>
<td>% yes</td>
<td>78.3</td>
<td>54.3</td>
<td>58.7</td>
<td>89.1</td>
<td>50.0</td>
<td>8.7</td>
</tr>
<tr>
<td><strong>Help Seeking</strong></td>
<td>57</td>
<td>% yes</td>
<td>77.2</td>
<td>50.9</td>
<td>59.6</td>
<td>89.5</td>
<td>50.9</td>
<td>8.8</td>
</tr>
</tbody>
</table>

Note. m coeff = number of alpha coefficients gleaned from journal articles; % yes = percentage of coefficients with matching samples coded as yes (1) for the predictor variable; % missing = percentage of coefficients with missing data for the coded variable. Only missing data percentages are presented for percent male.
Table 5. Results of OLS Regression of Moderator Variables for MSLQ Subscales.

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Intrinsic (m = 87)</th>
<th>Extrinsic (m = 60)</th>
<th>Task Value (m = 82)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(b)</td>
<td>SE</td>
<td>exp((b))</td>
</tr>
<tr>
<td>Intrinsic (m = 87)</td>
<td>−1.622</td>
<td>0.074</td>
<td>0.197</td>
</tr>
<tr>
<td>Extrinsic (m = 60)</td>
<td>0.012</td>
<td>0.035</td>
<td>1.012</td>
</tr>
<tr>
<td>Original item wording</td>
<td>0.067</td>
<td>0.040</td>
<td>1.069</td>
</tr>
<tr>
<td>Percent male</td>
<td>0.002</td>
<td>0.001</td>
<td>1.002</td>
</tr>
<tr>
<td>English</td>
<td>−0.240</td>
<td>0.077</td>
<td>0.787</td>
</tr>
<tr>
<td>Postsecondary</td>
<td>0.498</td>
<td>0.053</td>
<td>1.646</td>
</tr>
<tr>
<td>North America</td>
<td>−0.080</td>
<td>0.072</td>
<td>0.923</td>
</tr>
</tbody>
</table>

Control of Learning Beliefs \(m = 52\) | Self-Efficacy \(m = 157\) | Test Anxiety \(m = 59\)

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Intercept</th>
<th>SE</th>
<th>exp((b))</th>
<th>Intercept</th>
<th>SE</th>
<th>exp((b))</th>
<th>Intercept</th>
<th>SE</th>
<th>exp((b))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic (m = 87)</td>
<td>−0.752</td>
<td>0.108</td>
<td>0.471</td>
<td>−1.885</td>
<td>0.040</td>
<td>0.152</td>
<td>−1.309</td>
<td>0.076</td>
<td>0.270</td>
</tr>
<tr>
<td>Extrinsic (m = 60)</td>
<td>−0.098</td>
<td>0.052</td>
<td>0.907</td>
<td>−0.163</td>
<td>0.023</td>
<td>0.849</td>
<td>0.017</td>
<td>0.037</td>
<td>1.017</td>
</tr>
<tr>
<td>Original item wording</td>
<td>−0.258</td>
<td>0.062</td>
<td>0.773</td>
<td>−0.278</td>
<td>0.026</td>
<td>0.757</td>
<td>0.219</td>
<td>0.045</td>
<td>1.245</td>
</tr>
<tr>
<td>Percent male</td>
<td>−0.004</td>
<td>0.001</td>
<td>0.996</td>
<td>−0.001</td>
<td>0.001</td>
<td>0.999</td>
<td>0.002</td>
<td>0.001</td>
<td>1.002</td>
</tr>
<tr>
<td>English</td>
<td>−0.190</td>
<td>0.124</td>
<td>0.717</td>
<td>0.043</td>
<td>0.029</td>
<td>1.044</td>
<td>−0.503</td>
<td>0.057</td>
<td>0.605</td>
</tr>
<tr>
<td>Postsecondary</td>
<td>0.017</td>
<td>0.076</td>
<td>1.017</td>
<td>−0.062</td>
<td>0.031</td>
<td>0.940</td>
<td>−0.098</td>
<td>0.042</td>
<td>0.907</td>
</tr>
<tr>
<td>North America</td>
<td>0.222</td>
<td>0.097</td>
<td>1.249</td>
<td>−0.187</td>
<td>0.028</td>
<td>0.829</td>
<td>−0.114</td>
<td>0.048</td>
<td>0.892</td>
</tr>
</tbody>
</table>

Rehearsal \(m = 63\) | Elaboration \(m = 79\) | Organization \(m = 61\)

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Intercept</th>
<th>SE</th>
<th>exp((b))</th>
<th>Intercept</th>
<th>SE</th>
<th>exp((b))</th>
<th>Intercept</th>
<th>SE</th>
<th>exp((b))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Thinking (m = 67)</td>
<td>−1.118</td>
<td>0.060</td>
<td>0.327</td>
<td>−1.321</td>
<td>0.047</td>
<td>0.267</td>
<td>−1.420</td>
<td>0.061</td>
<td>0.242</td>
</tr>
<tr>
<td>Metacognition (m = 132)</td>
<td>−0.019</td>
<td>0.034</td>
<td>0.981</td>
<td>−0.070</td>
<td>0.027</td>
<td>0.933</td>
<td>0.065</td>
<td>0.033</td>
<td>1.067</td>
</tr>
<tr>
<td>Time Management (m = 57)</td>
<td>−0.066</td>
<td>0.048</td>
<td>0.936</td>
<td>0.097</td>
<td>0.037</td>
<td>1.102</td>
<td>0.029</td>
<td>0.047</td>
<td>1.030</td>
</tr>
<tr>
<td>Effort Regulation (m = 75)</td>
<td>−0.002</td>
<td>0.001</td>
<td>1.002</td>
<td>−0.002</td>
<td>0.001</td>
<td>0.998</td>
<td>0.001</td>
<td>0.001</td>
<td>1.001</td>
</tr>
<tr>
<td>Peer Learning (m = 37)</td>
<td>0.001</td>
<td>0.002</td>
<td>0.999</td>
<td>−0.006</td>
<td>0.001</td>
<td>0.994</td>
<td>0.003</td>
<td>0.001</td>
<td>1.003</td>
</tr>
<tr>
<td>Help Seeking (m = 46)</td>
<td>−0.009</td>
<td>0.001</td>
<td>0.999</td>
<td>−0.341</td>
<td>0.041</td>
<td>0.711</td>
<td>−0.387</td>
<td>0.088</td>
<td>0.679</td>
</tr>
</tbody>
</table>

Note. All variables except percent male are dichotomous categorical and coded 1 if study characteristics matched the descriptor and 0 if not. For example, if the study was conducted in North America, the article was coded 1, and coded 0 if not in North America. Outcome variable is bias-adjusted log-complement transformation of study alpha coefficient. Underlined values of \(b\) represent coefficients statistically significantly different from 0 at \(\alpha = .05\). \(b\) = estimate of regression coefficient; SE = standard error of \(b \times 10^2\); \(p\) = significance of \(t\) test for \(b\); exp(\(b\)) = back-transformation of \(b\); \(m\) = number of studies. OLS = ordinary least squares; MSLQ = Motivated Strategies for Learning Questionnaire.
alpha across the subscales. Although the English variable most consistently positively affected reliability, there was much more variability among the other moderators, which showed positive relationships with reliability for some subscales, but negative relationships for others. North America was denied credit in the model for four constructs in the OLS analysis; however, this variable was a significant moderator in the commonality analysis. Table 7 contains a summary of moderator effects for both OLS and commonality analyses.

**Seven-point response scales.** Positive and negative $b$ estimates of regression coefficients indicate that, for some subscales,

### Table 6. Commonality Analysis Results for MSLQ Subscales.

<table>
<thead>
<tr>
<th></th>
<th>Intrinsic ($m = 87$)</th>
<th>Extrinsic ($m = 60$)</th>
<th>Task Value ($m = 82$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unique</td>
<td>Common</td>
<td>Total</td>
</tr>
<tr>
<td>7-point response scale</td>
<td>0.01</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td>Original item wording</td>
<td>0.39</td>
<td>-0.01</td>
<td>0.38</td>
</tr>
<tr>
<td>Percent male</td>
<td>0.74</td>
<td>0.40</td>
<td>1.14</td>
</tr>
<tr>
<td>English</td>
<td>1.91</td>
<td>5.09</td>
<td>7.00</td>
</tr>
<tr>
<td>Postsecondary</td>
<td>20.76</td>
<td>5.53</td>
<td>26.29</td>
</tr>
<tr>
<td>North America</td>
<td>0.29</td>
<td>6.91</td>
<td>7.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Control of Learning Beliefs ($m = 52$)</th>
<th>Self-Efficacy ($m = 157$)</th>
<th>Test Anxiety ($m = 59$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unique</td>
<td>Common</td>
<td>Total</td>
</tr>
<tr>
<td>7-point response scale</td>
<td>1.33</td>
<td>-1.15</td>
<td>0.18</td>
</tr>
<tr>
<td>Original item wording</td>
<td>7.53</td>
<td>3.58</td>
<td>11.11</td>
</tr>
<tr>
<td>Percent male</td>
<td>5.38</td>
<td>-3.00</td>
<td>2.38</td>
</tr>
<tr>
<td>English</td>
<td>1.80</td>
<td>2.95</td>
<td>4.75</td>
</tr>
<tr>
<td>Postsecondary</td>
<td>0.03</td>
<td>0.38</td>
<td>0.41</td>
</tr>
<tr>
<td>North America</td>
<td>3.44</td>
<td>-2.30</td>
<td>1.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Rehearsal ($m = 63$)</th>
<th>Elaboration ($m = 79$)</th>
<th>Organization ($m = 61$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unique</td>
<td>Common</td>
<td>Total</td>
</tr>
<tr>
<td>7-point response scale</td>
<td>0.11</td>
<td>0.47</td>
<td>0.58</td>
</tr>
<tr>
<td>Original item wording</td>
<td>0.97</td>
<td>0.89</td>
<td>1.86</td>
</tr>
<tr>
<td>Percent male</td>
<td>1.38</td>
<td>0.99</td>
<td>2.37</td>
</tr>
<tr>
<td>English</td>
<td>0.02</td>
<td>0.83</td>
<td>0.85</td>
</tr>
<tr>
<td>Postsecondary</td>
<td>2.35</td>
<td>2.06</td>
<td>4.41</td>
</tr>
<tr>
<td>North America</td>
<td>0.25</td>
<td>0.30</td>
<td>0.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Critical Thinking ($m = 67$)</th>
<th>Metacognition ($m = 132$)</th>
<th>Time Management ($m = 57$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unique</td>
<td>Common</td>
<td>Total</td>
</tr>
<tr>
<td>7-point response scale</td>
<td>4.75</td>
<td>0.52</td>
<td>5.27</td>
</tr>
<tr>
<td>Original item wording</td>
<td>0.13</td>
<td>3.50</td>
<td>3.63</td>
</tr>
<tr>
<td>Percent male</td>
<td>0.30</td>
<td>0.11</td>
<td>0.41</td>
</tr>
<tr>
<td>English</td>
<td>0.00</td>
<td>6.41</td>
<td>6.41</td>
</tr>
<tr>
<td>Postsecondary</td>
<td>0.42</td>
<td>2.40</td>
<td>2.82</td>
</tr>
<tr>
<td>North America</td>
<td>1.11</td>
<td>6.27</td>
<td>7.38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Effort Regulation ($m = 75$)</th>
<th>Help Seeking ($m = 37$)</th>
<th>Peer Learning ($m = 46$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unique</td>
<td>Common</td>
<td>Total</td>
</tr>
<tr>
<td>7-point response scale</td>
<td>3.51</td>
<td>0.51</td>
<td>4.02</td>
</tr>
<tr>
<td>Original item wording</td>
<td>0.51</td>
<td>0.03</td>
<td>0.54</td>
</tr>
<tr>
<td>Percent male</td>
<td>0.02</td>
<td>0.33</td>
<td>0.35</td>
</tr>
<tr>
<td>English</td>
<td>1.43</td>
<td>0.07</td>
<td>1.50</td>
</tr>
<tr>
<td>Postsecondary</td>
<td>2.38</td>
<td>3.39</td>
<td>5.77</td>
</tr>
<tr>
<td>North America</td>
<td>0.12</td>
<td>0.30</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Note. $m =$ number of studies in each subscale analysis. MSLQ = Motivated Strategies for Learning Questionnaire.
7-point response scales positively influenced reliability, but in other subscales negatively influenced reliability (Table 5). For Task Value, Self-Efficacy, Elaboration, Critical Thinking, and Metacognitive Self-Regulation, 7-point response scales resulted in higher reliabilities than those with other point values. However, for Effort Regulation and Help Seeking, 7-point response scales resulted in lower reliabilities.

**Original item wording.** As seen in Table 5, original item wording was a statistically significant predictor of reliability scores for five subscales. However, commonality analysis results indicate original wording was also a predictor of reliability scores for Peer Learning and Time Management, and to a lesser extent for Organization and Critical Thinking. Regression coefficients, together with commonality analysis results, indicated that retaining the original wording tended to result in higher reliability coefficients for Self-Efficacy, Control of Learning Beliefs, Organization, and Critical Thinking.

**Postsecondary and percent male.** Postsecondary was a statistically significant predictor of reliability scores for Self-Efficacy, Test Anxiety, Rehearsal, Metacognition, and Peer Learning subscales. The positive regression coefficients suggest that when the sample consisted of postsecondary students, the reliability coefficients were lower for Intrinsic, Organization, and Effort Regulation. Although percent male was a statistically significant predictor of reliability scores for six factors, the regression coefficients were relatively close to zero on all those factors. The reason for such low $b$ values for this predictor is that variable values range from 0 to 100, whereas all other predictors are categorical and coded as either 0 or 1, which are on a similar scale as the criterion variable, transformed coefficient alpha. The negative regression coefficients indicated that when the sample consisted of more males, reliability scores were higher for Control of Learning Beliefs, Metacognition, and Elaboration.

**English and North America.** Commonality analysis results indicated that reviewing both unique and common contributions is important. Otherwise, the importance of North America would have been overlooked in three of the OLS results. North America was denied explanatory credit for Time Management and Help Seeking and to a lesser extent for Critical Thinking, although serving as suppressor effects for Metacognitive Self-Regulation. As seen in Table 7, of the 10 subscales in which North America served as a moderator, four demonstrated higher reliabilities when the study used North American participants. However, commonality analysis results also suggest three more scales would result in higher reliabilities using North American participants. Of the seven subscales in which English served as a moderator, six demonstrated higher reliabilities when the study was administered in English.

**Discussion**

One study conducted on one sample may provide information about a hypothesis, but improvements to a given field require comparisons of multiple studies with different study designs on various samples, and the ability to build on existing studies. Pintrich et al. (1991, 1993) understood that improvements to research on student motivation and learning would require the development of an instrument that was widely accessible to researchers. Thus, they developed the MSLQ. The result has been an instrument that has allowed
researchers to investigate motivation and learning across a variety of samples for many years. Methodologists have noted that score reliability varies across samples and does not relate to the reliability of tests (Thompson, 2003). In fact, researchers, editorial boards, and national educational and psychological associations have long recognized the importance of reporting reliability for the data in hand (Thompson, 1994; Wilkinson & APA Task Force on Statistical Inference, 1999). Determining how reliability estimates vary across administrations and samples can help future researchers as they plan their studies (Nimon et al., 2012). Therefore, we conducted an RG study to determine predictors of reliability scores for the MSLQ.

**RG of MSLQ Subscales**

If reliability of scores for a test is generalizable across studies, one would expect subsequent reliability estimates to remain consistent with those obtained during the development of instruments and published in original psychometric studies and test manuals. Vacha-Haase (1998) developed RG as a method to determine whether score reliability might be appropriately generalized across study populations with varying characteristics. For the present RG study, population estimates of alpha reliability for the 15 subscales of the MSLQ were generated using a newly applied VC technique (Bonett, 2010). Based on the disparity between estimates of mean reliability from the current study and reliabilities published in the MSLQ manual (Pintrich et al., 1991), and the presence of moderator variables that appear to influence reliability across studies, alpha reliabilities of the subscales of the MSLQ do not appear to generalize to various study populations.

The results indicate the extent to which variables predicted reliability scores varied across subscales. Original item wording was important for some subscales, but not others. Postsecondary samples tended to result in higher reliability for five subscales, lower reliabilities for three others, and did not serve as a predictor for the remaining subscales. The predictors in our model served as better moderators for Metacognitive Self-Regulation reliabilities than any other subscale.

**CIs for Mean Reliability Estimates**

CIs for mean reliability estimates in the current study for all 15 subscales did not encompass estimates reported in the MSLQ manual (Pintrich et al., 1991). The three lowest reliability estimates reported in the manual were below the lower bound of the CIs reported in Table 3. Therefore, a plausible range of score reliability for those factors (i.e., Extrinsic Goal Orientation, Organization, and Help Seeking) might be higher than the manual indicates, and Extrinsic Goal Orientation may fall within an acceptable range. However, for the remaining 12 subscales, reliability estimates for scores in future studies might actually be lower than those reported in the Pintrich et al. (1991) manual, although the plausible ranges often contain levels of reliability above .70.

Table 3 indicates that eight subscales have mean reliability CIs with a lower bound greater than .70, which is considered an acceptable level of score reliability, and one additional interval contains the value .70. CIs for mean reliability scores suggest Peer Learning might be lower than the .76 reported by Pintrich et al. (1991) and not at an acceptable level, but the larger standard error suggests that these reliabilities might vary more across samples than many of the other constructs. Organization, Rehearsal, and Control Beliefs have upper bounds close to an acceptable range, and the remaining subscales have reliability scores that are acceptable across a variety of samples. Although it is important to keep in mind the varying number of items in each of these subscales (see Table 1), no clear patterns or explanations are discernable from our data.

**Moderator Variables**

Of the 90 regression coefficients generated through OLS analysis, 39 were statistically significant at $\alpha = .05$, and for seven of the subscales, three or more of the six coefficients were statistically significant. Across all subscales in the study, variables that most commonly appeared to moderate reliability were postsecondary (eight subscales), 7-point scale (seven subscales), and English (seven subscales). Interestingly, none of the predictor variables appeared to moderate the Extrinsic Goal Orientation subscale, which may be an indication that reliability may generalize across a range of study factors for this scale.

Due to the log-complement transformation of study reliability in the regression analysis, estimates of $b$ are difficult to interpret. However, $\exp(b)$ may be interpreted as the percentage change in the nonreliability (i.e., $1 - \alpha$) for every 1-point change in the predictor variable (Bonett, 2010). A value of $\exp(b)$ greater than one represents a decrease in reliability. Interpretation of values of $\exp(b)$ for postsecondary suggests that studies utilizing the MSLQ with college students produce scores with significantly higher levels of reliability than younger students for the Self-Efficacy, Test Anxiety, Rehearsal, Metacognitive Self-Regulation, and Peer Learning subscales. However, significantly lower score reliability was indicated for postsecondary for the Intrinsic Goal Orientation, Organization, and Effort Regulation subscales (Table 5).

**Commonality Analysis**

Inspection of effect sizes presented in Table 6 indicates that the amount of variance of the reliability scores explained by the six predictor variables was noteworthy for Peer Learning ($R^2 = .4841$), Intrinsic ($R^2 = .3887$), Test Anxiety ($R^2 = .3581$), Time Management ($R^2 = .3591$), and Task Value ($R^2 = .3591$).
Unique and common contributions. The largest unique contributions were for postsecondary for the Intrinsic, English for Test Anxiety, and English for Time Management subscales. When multicollinearity exists, predictors might be denied explanatory credit, particularly if variables do not make a unique contribution but share variance with other predictor variables. Original item wording and North America were denied predictive credit in OLS regressions for both Time Management and Peer Learning. Denial of credit was due to minimal unique contributions these variables made to the models despite relatively high levels of shared variance with other predictors. Thus, if statistical significance of OLS regression coefficients was used as the sole criterion for making a contribution to the model, original item wording and North America would have been denied predictive credit. For this reason, it is recommended that use of multiple regression models to detect potential moderators in RG studies are followed with commonality analysis to identify cases where credit may be denied due to shared variance among the predictors.

Possible suppressor effects. English and North America possibly serve as suppressor effects for Metacognitive Self-Regulation. In addition, original item wording possibly serves as a suppressor for Help Seeking. Suppressor effects have indirect predictive power but improve the overall model (Burdenski, 2000; Courville & Thompson, 2001; Pedhazur, 1997; Thompson, 2006). In the suppressor case, the bivariate correlation between the suppressor variable and the dependent variable is close to zero, the beta weight is not close to zero, and inclusion of the suppressor in the model increases the effect size, $R^2$. Thus, researchers need to look beyond bivariate correlations when considering variables to include in future studies, otherwise important suppressor variables might be overlooked.

Study Limitations

As with any meta-analysis, the current study is limited by the quality and transparency of reporting in the included primary studies. We found that only 51% of studies reporting use of the MSLQ included reliability estimates for the data collected. The omission of studies may represent a publication bias, whereby authors may not report low score reliabilities, which potentially influences estimates of average reliability and the impact of moderator variables. Additional bias may also have been introduced by excluding dissertations, theses, and other gray literature, as authors of this research may be more likely to report unacceptably low reliability estimates than in peer-reviewed journals. In addition, the lack of reporting of key sample and study characteristics in primary studies served to limit both the sample size and the potential moderator variables available for analysis, which may result in underpowered moderator analyses (Hedges & Pigott, 2004) and model misspecification (Vacha-Haase & Thompson, 2011).

An additional limitation of the study is that the measure of reliability assessed was coefficient alpha, as that was the predominant internal consistency estimate reported. Although coefficient alpha assumes tau equivalence of factor loadings (Graham, 2006), a scant number of studies reporting on MSLQ data reported testing the assumption of tau equivalence. Uniquely, Berger and Karabenick (2011) reported using Raykov’s rho in lieu of coefficient alpha, because rho does not assume tau equivalence (Raykov, 1997).

Implications

The wide variability of alpha reported for MSLQ subscales suggests that researchers should not assume that future use will result in reliabilities similar to those reported in the test manual (Pintrich et al., 1991), especially when applying the instrument to populations and study conditions vastly different than the original norming study. Translations of the MSLQ typically resulted in decreased reliability estimates; thus, it is recommended that researchers requiring use of the MSLQ in languages other than English utilize established cross-cultural adaptation procedures, including back translation, cross-validation, and factor analysis (Sousa & Rojjanasrirat, 2011), or apply a previously validated adaptation (e.g., Lee, Yin, & Zhang, 2010). Researchers may also consider performing a small-scale RG on a carefully selected group of existing studies with characteristics similar to those in a planned future study to estimate expected reliabilities (Bonett, 2010). In addition, researchers reporting coefficient alpha should test for the assumption of tau equivalence or report a measure of reliability that does not assume invariant factor loadings. Potentially, a study could be conducted to determine whether measurement parameters are equal across groups reflected in the moderator variables to inspect item performance differences (cf. Vassar & Bradley, 2010).

Conclusion

Our findings provide further evidence of the importance of reporting score reliabilities rather than inducting reliability from other publications. Varying characteristics of the sample population, such as being in North America, can affect score reliability of nearly all subscales of the MSLQ. Similarly, modifying the MSLQ instrument, such as changing the 7-point scale to a 5-point scale or translating the instrument from English into another language, might positively or negatively affect subscale reliability.
When designing a quantitative study, forethought is important to administering tests that will result in reliability coefficients that are sufficient to produce unattenuated effect sizes (Pedhazur, 1997; Thompson, 2003; Yetkiner & Thompson, 2010). Reliability estimates obtained for an instrument will vary between applications if the score variability, sample composition, and administration conditions fluctuate (Crocker & Algina, 1986). Vacha-Haase and Thompson (2011) stated that

random variations in data, including the random variations associated with measurement error, attenuate the relationships among measured variables. Such attenuation occurs because correlation coefficients are sensitive to systematic covariances among measured variables replicated over study participants and not random fluctuations. (p. 159)

However, in certain circumstances, the attenuation of effect sizes due to low score reliability may not always occur (Nimon et al., 2012). Therefore, it is considered good practice to account for sample reliability in studies requiring statistical analysis and interpretation of data generated through the use of an instrument, such as in establishing test norms, assessing individuals and groups, performing validity studies, and evaluating sensitivity of measures. A failure to do so might have negative consequences for study outcomes and for individuals. For example, in clinical settings, use of inaccurate reliability estimates may result in misdiagnosis or underassessment, and in research settings, “insufficient reliability reporting practices influence the interpretation and application of research results and contribute to development and use of faulty measures” (Green, Chen, Helms, & Henze, 2011, p. 660).

Conceptually, reliability estimates the degree to which an individual’s scores remain relatively constant or free from measurement error over repeated administrations of the same test or of alternate forms of a test (Crocker & Algina, 1986; Rudner, 1994). Thus, to make a validity judgment about how well scores from an instrument measure a construct, researchers must consider score reliability. In this way, reliability is considered to be a necessary, but not sufficient condition for the establishment of score validity (Pedhazur & Schmelkin, 1991).

Regarding the negative implications of poor reliability reporting, it is common for researchers either to fail to report reliability estimates for data collected or to report only previously published reliability coefficients for the instrument, a practice that has been characterized as reliability induction (Vacha-Haase, Kogan, & Thompson, 2000). Such practice likely arises from a misunderstanding that reliability is a property of the scores generated by administration of a test to a particular sample under specific conditions, rather than a property of the instrument itself (Thompson & Vacha-Haase, 2000).

To encourage better practice in reliability reporting, journal editors (e.g., Thompson, 1994) and professional organizations (e.g., American Educational Research Association, 2006; APA, 2001; Wilkinson & APA Task Force on Statistical Inference, 1999) have consistently advocated that authors provide reliability estimates for the data in hand because “it is poor practice and potentially harmful to tested subgroups for researchers to assume that reliability evidence obtained with one sample (e.g., adult men) can generalize to other samples and/or populations (e.g., women, children, adolescents)” (Green et al., 2011, p. 658). Despite these efforts, reliability reporting practices appreciably have not improved over the past three decades (Green et al., 2011; Hogan, Benjamin, & Brezinski, 2000; Vacha-Haase & Thompson, 2011), an observation that is supported by findings from the current study.

Future researchers should benefit from the results of the current study as they allow for the estimation of reliability based on anticipated sample characteristics and study designs. In other words, researchers can better anticipate how their study design characteristics will affect the reliability of their results and make informed decisions about whether or not a particular modification (e.g., modification of item wording or use of a Likert-type scale with a different number of choices) is appropriate in light of the potential influence on score reliability. With these data in mind, researchers can make empirically based decisions to strengthen their research, and ultimately, the quality of scientific knowledge obtained using this measurement tool.
Appendix A

SPSS Code to Perform Ordinary Least Squares (OLS) Regression Moderator Analysis Using the Bonett Method

matrix.

***** Read data matrices from active dataset *****.

** X is an m x t matrix of predictors (m = studies, t = predictors)**. 
get X
   / variables = Xones to Xt
   / names = Xnames .

** P is an m x 1 vector of log-complement, bias-adjusted alpha estimates **. 
get P 
   / variables = P .

** V is an m x 1 vector of study sampling errors **. 
get V
   / variables = V .

***** calculate b hat *****.

calculate Xtrans = T(X) .
calculate XtransX = Xtrans * X .
calculate invXX = inv(XtransX) .
calculate bhat = invXX * Xtrans * P .

print bhat
   / title = “b Hat”
   / rnames = Xnames .

***** calculate cov(bhat) *****.

calculate Vdiag = mdiag(V) .
calculate XtransV = Xtrans * Vdiag .
calculate XtransVX = XtransV * X .
calculate covb = invXX * XtransV * invXX .
calculate varb = diag(covb) .

print varb
   / title = “b Variance”
   / rnames = Xnames .
end matrix .
Appendix B

Studies Included in the Reliability Generalization Meta-Analysis


Hsieh, P. H., Sullivan, J. R., Sass, D. A., & Guerra, N. S. (2012). Undergraduate engineering students’ beliefs, coping strategies,


strategies related to the affective component influence cognitive and metacognitive strategies. *Educación XXI*, 16, 231-246. doi:10.5944/eduexx1.16.2.2641


**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**References**


multiple regression case: An introduction to the package and a practical example. Behavior Research Methods, 40, 457-466. doi:10.3758/BRM.40.2.457


Romano, J. L., & Kromrey, J. D. (2009). What are the consequences if the assumption of independent observations is violated in reliability generalization meta-analysis studies? Educational and Psychological Measurement, 69, 404-428. doi:10.1177/0013164408323327


Yetkener, Z. E., & Thompson, B. (2010). Demonstration of how score reliability is integrated into SEM and how reliability affects all statistical analyses. Multiple Linear Regression Viewpoints, 36(2), 1-12.


Author Biographies

David F. Holland serves as the director of Planning, Assessment and Research in Birdville Independent School District.

Amanda Kraha is currently an assistant professor of Psychology at Indiana University East. She teaches undergraduate research methods, statistics, and cognitive psychology in both face-to-face and online environments. Kraha’s research centers on online teaching and learning, statistical techniques, and professional issues in psychology.
Linda R. Zientek is an associate professor of mathematics education at Sam Houston State University, Department of Mathematics and Statistics. Her research focuses on the preparation of mathematics teachers, developmental mathematics, and quantitative research methods.

Kim Nimon is an associate professor of Human Resource Development at The University of Texas at Tyler. Her research focuses on the workforce development and quantitative research methods.

Julia A. Fulmore is an assistant professor of Management at the University of Dallas, Satish & Yasmin Gupta College of Business. Her research interests are organizational commitment, organizational ethics, organizational culture, leadership, and quantitative research methods.

Ursula Y. Johnson is an assistant professor of pediatrics at the University of Texas Health Science Center at Houston, McGovern Medical School. Her primary research interests are developing and examining preventative interventions for young children at-risk for developmental delays or school failure.

Hector F. Ponce is an associate professor at the Universidad Autonoma de Ciudad Juarez. He teaches classes in quantitative research methods at the undergraduate and graduate programs, such as education, special education, educational research, and economics. His research focus is on psychometrics and statistics.

Mariya Gavrilova Aguilar teaches courses in human resource management and general business at the University of North Texas. She has expertise in performance improvement, instructional design, and career development. Her experience includes working in local government, casino, and university settings. Mariya conducts research in organizational behavior, human resource development, and teaching methods in management education.

Robin K. Henson is a professor and university distinguished teaching professor at the University of North Texas, Department of Educational Psychology. His teaching and research focus on applied quantitative methodology and measurement.