



www.ijemst.net

Reliability Generalization Meta-analyses in Mathematics Education Research: A Research Synthesis

Ashley M. Williams 
Texas A&M University, United States

Jamaal Young 
Texas A&M University, United States

To cite this article:

Williams, A. M. & Young, J. (2021). Reliability generalization meta-analyses in mathematics education research: A research synthesis. *International Journal of Education in Mathematics, Science, and Technology (IJEMST)*, 9(4), 741-759. <https://doi.org/10.46328/ijemst.1434>

The International Journal of Education in Mathematics, Science, and Technology (IJEMST) is a peer-reviewed scholarly online journal. This article may be used for research, teaching, and private study purposes. Authors alone are responsible for the contents of their articles. The journal owns the copyright of the articles. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of the research material. All authors are requested to disclose any actual or potential conflict of interest including any financial, personal or other relationships with other people or organizations regarding the submitted work.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.



International Journal of Education in Mathematics, Science, and Technology (IJEMST) affiliated with
[International Society for Technology, Education, and Science \(ISTES\): www.istes.org](http://www.istes.org)

Reliability Generalization Meta-analyses in Mathematics Education Research: A Research Synthesis

Ashley M. Williams, Jamaal Young

Article Info

Article History

Received:
24 December 2020
Accepted:
25 August 2021

Keywords

Reliability generalization
Meta-analysis
Mathematics education
STEM education
Instrumentation

Abstract

The purpose of this systematic review was to characterize the implementation of reliability generalization meta-analytic (RGM) practices within mathematics education-related empirical research. RGM studies are used to investigate and generalize the reliability of a measure across various studies. An exhaustive literature search was conducted to locate studies related to mathematics education, including RGM studies of psychological tests. The literature search included articles as well as grey literature (e.g., conference proceedings, dissertations, theses). Of the 27 RGM studies examined, five were on scales that related to mathematics education research, five were on scales related to motivation and/or learning, four related to self-esteem, self-concept, and/or self-efficacy, six related to perceptions, well-being, and/or anxiety, and seven related to personality or behavior. Of the mathematics education-related RGM studies, 85.5% (N=9,184) of the articles examined across studies had no mention of reliability or fell into the convention of citing previously reported reliabilities. Increasing awareness of RGM studies could lead to an increase in RGM studies conducted on mathematics education research scales, leading to increased understanding of mathematics education scales. This paper contributes to the literature on the practical and empirical importance of RGM for mathematics and STEM education praxis.

Introduction

Reliability remains an important consideration for mathematics educators seeking to improve their practice through research-informed decision making. The misapplication of the term reliability in educational and psychological research prevails, as evidenced by trends in researchers' reporting practices. Far too often, in various settings including academia, research, and clinical practice, authors incorrectly state that an instrument is reliable (King et al., 2014), which has led to the misconception that reliability is a property of a test or measure when it is truly a property of the scores produced by that instrument (Thompson & Vacha-Haase, 2000). This fallacy has direct implications on the effectiveness of interventions to improve the teaching and learning of mathematics.

Reliability illustrates to what extent scores yielded by an instrument administered to a target population, at a particular time, and under certain conditions are consistent and reproducible (Crocker & Algina, 1986; Onwuegbuzie & Larry, 2000; Taylor, 2012). Therefore, the reliability of scores produced by an instrument can be influenced by study and sample characteristics. Unfortunately, these score characteristics are rarely considered when researchers make conclusions and judgments that inform mathematics instructional praxis.

Traditional interpretations of reliability situate the test as the sole consideration upon which subsequent judgments are made. As a consequence of this false presumption that tests are reliable, researchers have overlooked the relevance of a correct interpretation of reliability (Cousin & Henson, 2000). Therefore, researchers fail to provide reliability estimates for data collected (Holland, 2015) or simply cite previously published reliability estimates, a practice known as reliability induction (Vacha-Haase et al., 2000).

Reliability estimates will vary across administrations (Crocker & Algina, 1986; Vacha-Haase et al., 2002; Vacha-Haase & Thompson, 2011) because anything that could potentially affect scores could also affect reliability (Barnes et al., 2002). Given the diversity across studies and the importance of score reliability in all quantitative analyses, it is pertinent that authors report reliability coefficients for their data (Vacha-Haase, 1998; Vacha-Haase & Thompson, 2011). Aside from the empirical importance of reliability reporting, there are equally important practical considerations.

The reliability of study outcomes is directly related to the efficacy of interventions and subsequent instructional decisions based on these outcomes. Improving score reliability reporting practices is imperative because study results can be influenced by reliability in various ways, potentially leading to misguided conclusions (Cousin & Henson, 2000; Greco et al., 2018). These considerations fall into three categories of challenges that arise from poor reliability reporting:

- First, reliability is a required condition for validity (Thompson, 2002); an instrument cannot be valid without first being reliable.
- Second, poor score reliability weakens the groundwork of frequently applied statistical analyses (Vacha-Haase & Thompson, 2011), leading to decreased estimates of statistical significance and effect sizes (Greco et al., 2018; Thompson, 2002; Yetkiner & Thompson, 2010).
- Lastly, failure to report reliability as it pertains to the particular study and sample characteristics compromises your study's replicability even under similar conditions (Cousin & Henson, 2000). Failure to report reliability could lead to negative consequences for individuals and study outcomes (Holland, 2015) and weaken the empirical quality of present and future research.

Over two decades ago, to emphasize the importance of score reliability and encourage authors to report reliability coefficients for their data, a new methodological approach to explore reliability coefficients across studies emerged. This new approach, called reliability generalization (RG) or reliability generalization meta-analysis (RGM), described by Vacha-Hasse (1998), provided a means for illustrating score integrity and characterizing study features that may predict variations in score quality. Although the meta-analysis of reliability coefficients has been around for decades (e.g., Jacoby & Matell, 1971; Lissitz & Green, 1975;

Churchill & Peter, 1984), the term RG wasn't used until the early '90s (e.g., Kennedy & Turnage, 1991) and didn't become popular until Vacha-Haase (1998) proposed it as an extension to the already existing meta-analytic method of Validity Generalization developed by Schmidt and Hunter (1977) and Hunter and Schmidt (1990). She set the foundation for reliability generalization with her psychometric meta-analysis of reliability coefficients reported for the Bem Sex-Role Inventory (Vacha-Haase, 1998). Following Vacha-Haase's seminal study, dozens of RGM studies have been published (Vacha-Haase & Thompson, 2011). However, a synthesis of the outcomes of RGMs concerning mathematics education remains elusive.

RGM provides researchers with a method to assess the score reliability in prior applications of an instrument and investigate possible sources of reliability estimates variability. Insight of this nature may help guide researchers' plans of future studies, estimate anticipated levels of reliability, and advise study design choices concerning effect sizes, power, and statistical significance (Henson & Thompson, 2002). Thus, RGM is a promising methodological innovation to support increased reliability across mathematics education research.

The benefits of RGM are threefold. First, RGM leads to a deeper understanding of various instruments (Cousin & Henson, 2000), and results obtained can suggest populations, samples, or groups for which particular instruments may be more or less appropriate, providing evidence that can aid researchers in improving instruments or adapting the instrument for dissimilar populations of interest (Taylor, 2012). Hess, McNab, & Basoglu (2014) suggest that RGM is an important step towards holistic evaluations of construct validity, affording insight on how certain study characteristics can reduce scale validity or even render a scale unsuitable for some settings.

Additionally, when describing instrument selection and application, having an available RGM would provide comparative data to facilitate interpretations of outcomes (Leech et al., 2011). As more RGM publications arise, it could potentially stimulate more comprehensive discussions of score reliability in the literature (Cousin & Henson, 2000), hence stressing the importance of reliability and encourage authors to report score reliability for their data. Most importantly, RGM findings confront the engrained misconception that reliability is a property of the test and communicates the importance of understanding that score reliabilities are sample dependent and often vary across administrations (Vacha-Haase & Thompson, 2011). In the next section, we outline the organization and structure of the present study.

In the sections that follow, we first situate RGM as a possible means of addressing some of the enduring challenges within mathematics education, situating the problem in the proper context. Based on the promise of RGM to guide future investigations within mathematics education, we then present the purpose of the present study and the potential implications for mathematics teaching, learning, and future research. Then we review the literature to summarize prior approaches to RGM and their outcomes across other disciplines. Next, we describe the research methods and data analysis procedures used to collect and evaluate the present study's data. Finally, we present the results of the current study and provide recommendations to support research and instructional practice within mathematics education.

Problem Statement

RGM studies are used to investigate and generalize the reliability of a measure across various studies. RGM studies also illustrate the variability of score reliabilities and establish in what circumstances the score reliability may be unacceptable (Caruso, 2000; Vacha-Hasse, 1998; Vacha-Haase et al., 2002; Vacha-Haase & Thompson, 2011). In RGM, one primary goal is to identify the source of measurement error across studies using the same instrument. To address this goal, studies become the unit of analysis, the reliability coefficients become the dependent variables, and scale, study, or sample characteristics become possible predictors (Cousin & Henson, 2000). Ultimately, determining these sources of error helps researchers make valuable decisions such as selecting a target population given a particular instrument or inversely choosing an appropriate instrument given your population of interest, which may lead to more precise interpretations and conclusions of results.

Increasing the next generation of adults' mathematics literacy is an international challenge, as opportunities to learn, declining interest in mathematics, and gender gaps abound. Researchers and teachers need effective and efficient instructional resources to address these and other challenges within mathematics education. Better instructional resources cannot be realized until the reliability of study data is systematically addressed within mathematics education. RGM is one means to initiate conversations and action within mathematics education related to the importance of reliability generalization within mathematics education research.

Aim of the Study

This systematic review aimed to characterize the implementation of reliability RGM practices within mathematics education-related empirical research. RGM studies are used to investigate and generalize the reliability of a measure across various studies. Through the present systematic review, we hope to summarize current and prior approaches to RGM as a means to impact future studies to support the efficacy of interventions in mathematics education by improving the reliability of study outcomes. In the sections below, we illustrate the rationale and results of this research endeavor, but first, we review the prior methodological approaches to RGM related to the present project.

Literature Review

The application of RGM is well-documented but varies across the research literature. The transformation of alpha is often recommended to adhere to the assumptions of many tests within a meta-analysis. Several RGM studies analyze untransformed coefficients alpha (e.g., Vacha-Haase, Tani et al., 2001; Leach et al., 2006) as recommended by some authors (e.g., Henson & Thompson, 2002; Thompson & Vacha-Haase, 2000). Others express a concern that the Cronbach's alpha is both bounded and not normally distributed; thus, Cronbach's alpha estimates violate the assumption that effect sizes are normally distributed in a meta-analysis (Feldt & Charter, 2006; Rodriguez & Maeda, 2006; Shou & Olney, 2020). Therefore, before modeling, a transformation of the alpha coefficients is necessary. The Hakstian-Whalen (1976), Fisher's r to Z , and Bonett transformation (Bonett, 2002) are all transformation methods that are used in the RGM literature (Semma et al., 2019).

Markedly, Fisher's r -to- Z transformation is not recommended for the transformation of alpha (Feldt & Brennan, 1989; Henson & Thompson, 2002). Sánchez-Meca et al. (2012) indicates that based on a simulation study, under the homogeneity assumption, the transformation of coefficient alpha had very little influence on the average coefficient alpha. Nonetheless, the preparation of alpha coefficients is an essential consideration in an RGM study.

Meta-regression models are well represented within RGM research. There are four common regression models used to integrate a set of alpha coefficients: the fixed-effects (FE) model (Hedges & Olkin, 1985), the random-effects (RE) model (Hedges & Vevea, 1998), the varying-coefficient model (Bonett, 2010; Laird & Mosteller, 1990), and the mixed-effects model (Raudenbush & Bryk, 1985). The fixed-effects and varying-coefficient models are used to generalize results only to studies with similar characteristics to those included in the meta-analysis, where random-effects is generalizable to a broader population of studies (Sánchez-Meca et al., 2012).

The different statistical methods vary depending on the need to transform or weight the reliability coefficients. Transformation of coefficient alpha is required for fixed-effects and varying-coefficients models and highly advised for random-effects models; fixed-effects and random-effects models provided weighted mean alpha coefficients, while varying-coefficients provides a simple arithmetic mean (Sánchez-Meca et al., 2012). Akin to general regression, meta-regression is a powerful and popular approach used by researchers conducting RGM studies.

Applying representative weights is appropriate across the spectrum of meta-analytic approaches. There are four weighting methods applied in RGM studies. The first RGM method documented in the literature (e.g., Vacha-Haase, 1998) is to work with the unweighted coefficient alpha applying the ordinary least squares (OLS) technique where an average coefficient alpha is obtained by calculating the simple arithmetic mean of the untransformed reliability estimate. The second and third method is weighting by inverse variance for fixed-effects models and random-effects models; additional weighting for random-effects models accounts for both within and between-study variance, which tends to produce wider confidence intervals than fixed-effects models (Holland et al., 2018; Sánchez-Meca et al., 2012). Lastly, weighting untransformed coefficients alpha by sample size gives more weight to the studies with larger sample sizes (Yin & Fan, 2000). Weighting schemes can be found on page 406 of Sánchez-Meca et al. (2012). Deciding on a weighting scheme will influence the statistical method you choose; random-effects and fixed-effects models will apply a weighting scheme, whereas if an unweighted mean is desired other statistical models, including the varying-coefficients model, will have to be applied. The application of weighting schemes is thus an important component of RGM.

As in most meta-analyses, the assessment of moderators is arguably just as important as the effect sizes summary statistics. There are various approaches applied in RGM research to assess moderator variables and explain reliability coefficients variability. Some researchers suggest several descriptive and inferential statistical strategies to explore variability (e.g., Vacha-Haase, Tani et al., 2001). This technique was often applied when an insufficient sample size didn't allow for regression (e.g., Capraro & Capraro, 2002). Others that faced this issue used correlational analyses (e.g., Henson et al., 2001; Nilsson et al., 2002). Often researcher used a combination

of ANOVA to analyze categorical predictors and multiple regression for continuous predictors (e.g., Aguayo et al., 2011; Lane et al., 2002; Taylor, 2012), or correlational analyses paired with multiple regression (Mahapoonyanont et al., 2010; Vassar et al., 2011).

Several researchers applied waves of multiple regression analysis for categorical and continuous predictors (Capraro et al., 2001; Caruso et al., 2001; Ross et al., 2005; Warne, 2011). Meta-regression models can also be used to analyze moderators such as the use of multiple RE analyses for each predictor (e.g., Leue & Lange, 2011), or a mixed-effects model (ME) which includes both fixed- and random-effects (Beretvas & Pastor, 2003) to combined all predictors into one model (e.g., Shou & Olney, 2020). When multiple levels of nested predictors exist, an HLM approach can be employed (Wang, 2002). Together, these methods represent the most frequently used approaches to moderator analysis present in educational literature.

Method

An exhaustive literature search was conducted to locate studies related to mathematics education, including RGM studies of psychological tests. The literature search included articles as well as grey literature (e.g., conference proceedings, dissertations, theses). Only studies meeting the following criteria were included in this systematic review:

- (a) author(s) conducted an RGM that presented summary statistics of score reliability across studies,
- (b) the study examined instruments used in prior mathematics education or related STEM education research,
- (c) the study was published between 2000 and 2020.

The initial search was conducted through the university library collection of databases, was limited to text in English, and used the search terms "*reliability generalization*" OR "*meta-analysis of reliability*" OR "*Psychometric meta-analysis*" OR "*Meta-Analysis of Coefficient Alpha*" AND "*mathematics*" OR "*math*" OR "*mathematics education*" OR "*math education*" OR "*STEM*." The database search resulted in 1015 hits. Medical or business-related databases were then excluded, resulting in 512 hits, which narrowed the pool to 356 studies once the search engine removed duplicates. During an inclusive scan of titles and abstracts, 44 studies were further inspected for exclusion, resulting in 13 studies that met the criteria.

The second search was conducted in Google Scholar. We applied the same search phrases and received 1950 hits. After scanning titles and information provided on results, this was narrowed to 54 hits. Further inspection resulted in 12 studies that met the criteria. Lastly, references of included papers were searched, and two additional studies were located for inclusion. The search and retrieval process resulted in 27 studies that were included in the present literature synthesis. We present the combined database and Google search process in Figure 1.

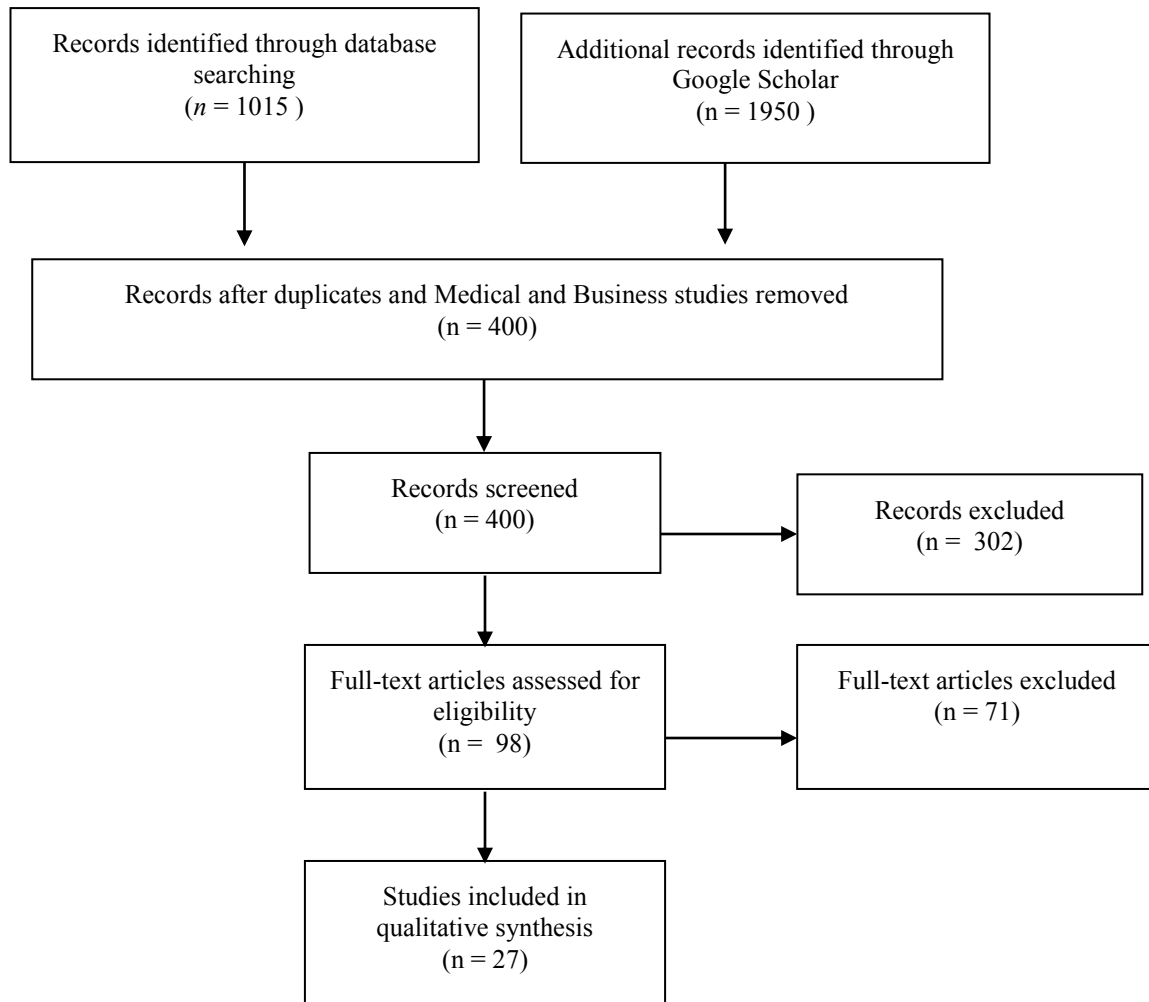


Figure 1. Outline of Literature Search and Retrieval Process

Findings and Discussion

Of the 27 RGM studies examined, five were on scales that related to mathematics education research, five were on scales related to motivation and/or learning, four related to self-esteem, self-concept, and/or self-efficacy, six related to perceptions, well-being, and/or anxiety, and seven related to personality or behavior. Two of the 27 studies used RGM techniques to analyze reliability within a large data set and were excluded from further investigation. The first excluded study utilized a data set consisting of a one-time administration across several schools (Maeda & Rodriguez, 2002). The second eliminated study examined the administration of a scale given to multiple combinations of groups (Kettler et al., 2011). A third study was excluded because researchers could not complete the meta-analysis due to insufficient data (Wilder & Sudweeks, 2003). Details on the remaining 24 RGM studies, including scales analyzed, general methods, and significant moderators, can be found in Table 1.

Table 1. Prior RGM in or Useful to Mathematics Education Research

Author(s)	Year	Type	Scale(s)	n_1	n_2	Methods	Moderators	ρ_{type}	ρ_{mean}
Capraro, Capraro, & Henson	2001	A	Mathematics Anxiety Rating Scale	7	50	Descriptive statistics; Multiple regression	Age, reliability type, number of items, total score SD	Overall Alpha T-RT	.900 .915 .841
Caruso, Witkiewitz, Belcourt- Dittloff, & Gotlieb	2001	A	Eysenck Personality Questionnaire Subscales: Psychoticism (P) Extraversion (E) Neuroticism (N) Lie (L)	52	1175	Compared untransformed, squared, & Fisher's z transformation Multiple regression	P: SD_{scores} , SD_{age} , sample type E: SD_{scores} N: SD_{scores} L: SD_{scores} , SD_{age}	Alpha PENL	.66 .83 .83 .77
Henson, Kogan, & Vacha-Haase	2001	A	Teacher Efficacy Scale subscales: Personal teaching efficacy (PTE) General teaching efficacy (GTE)	52	NR	Box plots; bivariate correlational analysis	Experience, level, area, gender, sample size, score variance, number of items	Alpha PTE GTE	.778 .696
			Science Teaching Efficacy Belief PSTE STOE				Experience, level, gender, sample size, score variance, number of items	Alpha PSTE STOE	.885 .761
			Teacher Locus of Control Subscales Student success I+ Student failure I-				Experience, level, sample size, score variance	Alpha I+ I-	.740 .700
			Responsibility for Student Achievement Student Failure RSA + Student Success RSA -				Area, gender, sample size, number of items	Alpha RSA+ RSA-	.760 .840
Nilsson, Schmidt, & Meek	2002	A	Career Decision- Making Self- Efficacy Scale	20	29	ANOVA, correlation	SD_{scores} , race, age	Alpha Full form Short form	.95 .94
Vacha-Haase, Kogan, Tani, & Woodall	2001	A	Minnesota multiphasic personality inventory 10 subscales (1) - (0)	153	1819	Mean, median, confidence intervals, box plots of untransformed & unweighted reliability estimates, multiple	(1) (2) (4) (5) (7) - adults vs adolescents (1) (3) - form, (2) nonclinical (3) vs. (4)	Alpha + T-RT (1) (2) (3) (4)	.72 .70 .65 .66 .67 .64

						regression	incarcerated	(5)	.72
							(6) – form,	(6)	.73
							adults vs.	(7)	.69
							adolescents	(8)	.81
							(8) (9) (0) –	(9)	
							adults vs	(0)	
							college		
							students		
Vacha-Haase, Tani, Kogan, & Woodall	2001	A	Minnesota multiphasic personality inventory MMPI/MMPI 2 validity subscales Lie (L) Infrequency (F) Correction (K)	153	1819	Box-and-whisker plots, descriptive statistics of untransformed & unweighted reliability estimates, multiple regression	L – reliability type, adult vs. nonadult, college vs. noncollege F – reliability type, college vs. noncollege, clinical vs. nonclinical K - adult vs. nonadult, college .vs noncollege	Alpha + T-RT L F K	.68 .68 .73
Barnes, Harp, & Jung	2002	A	Spielberger state-trait anxiety inventory state-Trait	46	770	Descriptive statistics, correlation analysis	Test form, age, SD _{scores}	Alpha State Trait T-RT State Trait	.91 .89 .70 .88
Mji & Alkhateeb	2005	A	Conceptions of mathematics questionnaire	8	NR	ANOVA		Alpha Fragmented Cohesive	.89 .91
Crouch	2016	T	Ryff's Scale of Psychological Well-Being	264	NR	RE model, ANOVA	Age, number of items, response format, the language e of test	Alpha	.858
Capraro & Capraro	2002	A	Myers-Briggs Type Indicator Extravert-introvert (EI) Sensing-intuitive (SN) Thinking-feeling (TF) Judgement- perception (JP)	14	196	Descriptive Statistics		Overall Alpha T-RT EI SN TF JP	.815 .816 .813 .838 .843 .764 .822
Henson & Hwang	2002	A	Kolb's Learning Style Inventory	34	81	Descriptive Statistic, multiple regression	Test form, academic	Alpha T-RT	

			subscales:			major, setting	CE	.75 ^m	
			Concrete					.40 ^m	
			experience (CE)				RO	.79 ^m	
			Reflective					.52 ^m	
			observation (RO)				AC	.80 ^m	
			Abstract					.56 ^m	
			conceptualization				AE	.81 ^m	
			(AC)					.55 ^m	
			Active						
			experimentation						
			(AE)						
Lane, White, & Henson	2002	A	Coopersmiths self-esteem inventory	33	244	Calculated estimates with KR-21 when applicable, ANOVA, multiple regression	Number of items, ethnicity, risk status, age, intelligence	Overall KR21/Alpha T-RT KR20	.644 .725 .538 .648
Vassar, Knaup, Hale, & Hale	2011	A	The Impact of Event Scales Composite	66	232	Descriptive statistics, correlation, multiple regression	War & abuse victims, % female, journal type	Alpha	.87
Ross, Blackburn, & Forbes	2005	A	Patterns of adaptive learning survey	30	11	Descriptive statistics, multiple regression	Scale version, cited manual date	Alpha Composite E TGO PAP PAV	.77 .68 .79 .79 .81
Leach, Henson, Odom, & Cagle	2006	A	Self-Description Questionnaire SDQ1 Total academic (TA) Total nonacademic (TNA) General self- concept (GSC) SDQ2 Math (M) Reading (R) General School (GS) General self- concept (GSC2)	56	56	Descriptive statistics (no transformation), Multiple regression, ANOVA	LS, test adjusted, school type, age, SES	Alpha SDQ1 TA TNA GSC M R GS GSC2	.92 .91 .88 .79 .92 .85 .87 .86
Mahapoonyan ont, Krahamwong, Kochakornjaru	2010	A	Robert H. Ennis's critical thinking concept	14	11	Fisher's z transformation, weighted, correlation, multiple regression	Multiple choice scale	Alpha	.897

pong, & Rachasong									
Aguayo, Varga, Fuente, & Lozano	2011	A	The Maslach Burnout Inventory Emotional Exhaustion (EE) Depersonalization (D) Personal accomplishment (PA)	45	9	T transformation, weighted by inverse sample variance, ANOVA, multiple regression	SD _{score} , country, age, sample type, test language, test version	Alpha EE D PA	.88 .71 .78
Leue & Lange	2011	A	Positive Affect (PA) and Negative Affect (NA) Schedule	109	139	measurement error correction, RE model (run multiple times), fail-safe	adults vs. adolescents, clinical vs. nonclinical, short term vs. long term, English vs. non-English	Alpha T-RT PA NA	.894 .586 .848 .569
Warne	2011	A	Overexcitabilit y Questionnaire– Two Subscales: Intellectual (I) Imaginational (IM) Emotional (E) Sensual (S) Psychomotor (P)	11	2	T-transformation calculated variances and weights for each FE model, multiple regression	I – sample variance, origin, age IM – sample variance, sample size, origin, #items, age E – sample variance, age S – sample variance, # items, age P – gender, sample variance, age	Alpha I IM E S P	.859 .850 .822 .871 .850
Taylor	2012	D	Motivated Strategies for Learning Questionnaire Intrinsic goal (IG) Extrinsic goal (EG) Task Value (TV) Control of learning beliefs (CB) Self-efficacy	123	102			Alpha IG EG TV CB SE TA R E O CT MSR TSM ER	.71 .68 .85 .65 .88 .76 .68 .76 .70 .79 .78 .73 .62 .68

			(SE)				PL	.61	
			Test anxiety				HS		
			(TA)						
			Rehearsal (R)						
			Elaboration						
			(E)						
			Organization						
			(O)						
			Critical						
			Thinking (CT)						
			Metacognitive						
			self-regulation						
			(MSR)						
			Time & study						
			management						
			(TSM)						
			Effort						
			regulation						
			(ER)						
			Peer Learning						
			(PL)						
			HHelp-						
			Seeking(HS)						
Hess, McNab, & Basoglu	2014	A	Perceived Ease of Use (PEOU), Perceived Usefulness (PU), & Behavioral intentions (BI)	347	NR	RE model, correlation, single moderator analysis, two-factor analysis main & interaction terms regression; multiple moderator model	PEOU - Tech purpose, reliability type, original scale items PU - tech purpose, experience, Language, original scale items BI - gender, language, reliability type	Alpha PEOU PU BI	.873 .888 .880
Shou & Olney	2020	A	domain-specific risk-taking (DOSPERT) scale Ethical (E) Financial (F) Health (H) Recreational (R) Social (S)	94	830	Bonnet-transformation, RE model, ME model	Rating aspect, Version, target population, sample type, mean age, language, gender	Alpha Total E F H R S	.87 .73 .78 .71 .80 .68
Kilgus, Eklund,	2018	A	Student Risk Screening	7	9	Bonnet transformation, weight studies by the		Alpha	.83

Maggin, Taylor, & Allen			Scale			inverse of the squared standard error of the sampling distribution, FE & RE model			
Holland et al.	2018	A	Motivated strategies for learning questionnaire	95	28	VC model, GLM and OLS multiple regression method	IG: language, setting	Alpha IG	.709 .692
			Intrinsic goal (IG)				TV: LS, gender, language, country	EG TV	.833 .645
			Extrinsic goal (EG)				CB: wording, gender, country	CB SE	.879 .759
			Task Value (TV)				SE: LS, wording, setting, country	TA R	.668 .745
			Control of learning beliefs (CB)				TA: wording, setting, country,	E O	.679 .778
			Self-efficacy (SE)				language	CT	.754
			Test anxiety (TA)				R: setting	MSR	.724
			Rehearsal (R)				E: LS, gender, wording, country	TSM ER	.660 .628
			Elaboration (E)				O: language, setting	PL HS	.608
			Organization (O)				CT: LS		
			Critical Thinking (CT)				MSR: LS, gender, language, country, educational setting		
			Metacognitive self- regulation (MSR)				TSM: gender, language		
			Time & study management (TSM)				ER: LS, educational setting		
			Effort regulation (ER)				PL: gender, setting		
			Peer Learning (PL)				HS: LS, wording,		
			HHelp-Seeking(HS)						

Notes. A = published article. D = dissertation. T = thesis. NR = not reported. PSTE = personal science teaching efficacy. STOE = Science teaching outcome expectancy. T-RT = test-retest. SD = standard deviation. LS = Likert scale $n1$ = number of articles that reported reliability. $n2$ = total number of articles that had no mention of reliability or merely cited previously reported reliability. ρ_{type} = type(s) of reliability estimates. ρ_{mean} = mean of reliability estimates. ^mMean reliability score was not provided replaced with median. Moderators included were only those found important or significant in explaining the variability of reliability coefficients.

During the examination of the 24 RGM studies, the underreporting of reliability mirrored prior research findings (e.g., Barnes et al., 2002; Caruso & Edwards, 2001; Vacha-Haase & Thompson, 2011). Of the mathematics education-related RGM studies, 85.5% ($N=9,184$) of the articles examined across studies had no mention of reliability or fell into the convention of citing previously reported reliabilities. Vacha-Haase and Thompson (2011) illustrated in their examination of RGM studies that 70.3% ($N=12,994$) of primary studies failed to mention or merely cited reliability estimates. Therefore, taking a closer look at the RGM studies that dated from 2012 to 2020 in the present systematic review, we observed that 70.1% ($N=1,826$) still had no mention of reliability or practiced reliability induction. The lack of reliability reporting exemplifies that poor reporting practices still exist. Continuing to conduct and advocate for the application of RGM could lead to a better understanding of reliability and improvement in reporting practice.

Concerning RGM methodology, model reporting was absent. Of the 24 included studies, 5 (20.83%) applied a random-effects model, 2 (8.33%) used a fixed-effects model, and the majority (17 or 70.83%) did not indicate which model was applied in the meta-analysis. Finally, the approaches applied to characterize RGM data varied but tended to represent more traditional approaches, such as multiple regression and ANOVA. Subsequently, HLM was not used in any of the mathematics education-related RGM studies. From the included studies, the majority (14 or 58.33%) used a multiple regression approach, 6 (25%) used ANOVA, and 2 (8.33%) used both box-and-whisker plots and bivariate correlation. In the discussion that follows, we unpack these results to illustrate how the findings can inform future research and mathematics education practice.

There are numerous benefits of RGM studies that can improve mathematics education research. Therefore, these data must be accessible and accurate. Synthesized assessments of scale reliability can inform researchers as they select an instrument for their population of interest (King et al., 2014). For example, understanding what types of instruments work best with certain age groups or in certain settings could lead researchers to determine what instruments would be best for the population they are interested in. Because we as educational researchers are responsible for our field and the communities we serve, we must use scales that are most appropriate to conduct research efficiently and effectively. Increasing awareness of RGM studies could lead to an increase in RGM studies conducted on mathematics education research scales, leading to increased understanding of mathematics education scales. The identification of study characteristics impacts the praxis of mathematics education. Mathematics education researchers could also harness this information to improve instruments and study designs. This knowledge can increase statistical power, influence and improve interpretations of results, which are an integral aspect of complete research reporting (Caruso et al., 2001), and strengthen instrument validity.

The results of the present study indicate that poor reporting practices abound within the observed studies. Poor reporting practices are problematic, as the results of research studies are only as good as the validity and the reliability of the instruments used to collect data. Here we noticed that the vast majority of studies did not report reliability coefficients or practiced reliability induction (i.e., reporting reliability scores from prior research). The absence of these data creates a sizable challenge for the utility of the findings, as the absence of these data limits the generation and possibility reduces the likelihood of study replication. Aside from a general lack of reporting of reliabilities within primary studies, there was also a lack of utilization of more advanced statistical techniques such as HLM, which help to account for the influence of the data structure, which is often an artifact of study designs. RGM researchers' over-reliance on ANOVA and meta-regression places unnecessary limitations on the questions that can be asked and subsequently answered by RGM by inhibiting the ability of the methodology and, more importantly, the field from moving forward.

Conclusion

In conclusion, information provided by RGM studies also can help bridge several gaps in mathematics education, such as achievement, opportunity, and equity. Mji and Alkhateeb (2005) suggest relating such psychometric properties of scores to better understand the effects of misconceptions on mathematics learning and understanding. Additionally, RGM studies of large data sets can influence standardized testing practices in the future (see Maeda & Rodriguez, 2002; Ryngala et al., 2005). Given the several important contributions

RGM studies can provide to the field, it is inexplicable why they are not more commonly utilized in mathematics education research. Raising awareness of the benefits of RGM could strengthen the empirical quality of mathematics education research.

References

(References with an asterisk indicate studies included in the analysis.)

- *Aguayo, R., Vargas, C., Emilia, I., & Lozano, L. M. (2011). A meta-analytic reliability generalization study of the Maslach Burnout Inventory. *International Journal of Clinical and Health Psychology, 11*(2), 343–361. <https://www.redalyc.org/pdf/337/33716996009.pdf>
- *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%2F0013164402062004005>
- 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%2F0013164402239318>
- 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%2F10769986027004335>
- Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological methods, 15*(4), 368–385. <https://psycnet.apa.org/doi/10.1037/a0020142>
- *Capraro, M. M., Capraro, R. M., & Henson, R. K. (2001). The measurement error of scores on the mathematics anxiety rating scale across studies. *Educational and Psychological Measurement, 61*(3), 373–386. <https://doi.org/10.1177%2F00131640121971266>
- Capraro, R. M., & Capraro, M. M. (2002). Myers-Briggs type indicator score reliability across Studies a meta-analytic reliability generalization study. *Educational and Psychological Measurement, 62*(4), 590–602. <https://doi.org/10.1177%2F0013164402062004004>
- Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement, 60*(2), 236–254. <https://doi.org/10.1177%2F00131640021970484>
- Caruso, J. C., & Edwards, S. (2001). Reliability generalization of the Junior Eysenck Personality Questionnaire. *Personality and Individual Differences, 31*(2), 173–184. [https://doi.org/10.1016/S0191-8869\(00\)00126-4](https://doi.org/10.1016/S0191-8869(00)00126-4)
- *Caruso, J. C., Witkiewitz, K., Belcourt-Dittloff, A., & Gottlieb, J. D. (2001). Reliability of scores from the Eysenck Personality Questionnaire: A reliability generalization study. *Educational and Psychological Measurement, 61*(4), 675–689. <https://doi.org/10.1177/00131640121971437>
- Churchill, G. A., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research, 21*(4), 360–374. <https://doi.org/10.1177%2F002224378402100402>
- Cousin, S. L., & Henson, R. K. (2000). *What is reliability generalization, and why is it important?* (ED445077). ERIC. <https://files.eric.ed.gov/fulltext/ED445077.pdf>

- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.
- *Crouch, M. (2016). *Reliability generalization: Exploring score reliability variance with Ryff's scale of psychological well-being* [Unpublished master's thesis]. Brock University.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105–146). New York: American Council on Education and Macmillan.
- Feldt, L. S., & Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement*, 66(2), 215–227. <https://doi.org/10.1177%2F0013164404273947>
- Greco, L. M., O'Boyle, E. H., Cockburn, B. S., & Yuan, Z. (2018). Meta- analysis of coefficient alpha: A reliability generalization study. *Journal of Management Studies*, 55(4), 583-618. <https://doi.org/10.1111/joms.12328>
- Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significant test for independent alpha coefficients. *Psychometrika*, 41(2), 219–231.
- Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. London: Academic Press.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, 3(4), 486–504.
- *Henson, R. K., & Hwang, D. Y. (2002). Variability and prediction of measurement error in Kolb's learning style inventory scores a reliability generalization study. *Educational and Psychological Measurement*, 62(4), 712–727. <https://doi.org/10.1177/0013164402062004011>
- *Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the teacher efficacy scale and related instruments. *Educational and Psychological Measurement*, 61(3), 404–420. <https://doi.org/10.1177/00131640121971284>
- 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>
- *Hess, T. J., McNab, A. L., & Basoglu, K. A. (2014). Reliability generalization of perceived ease of use, perceived usefulness, and behavioral intentions. *MIS Quarterly*, 38(1), 1–28. <https://doi.org/10.25300/MISQ/2014/38.1.01>
- Holland, D. F. (2015). *Reliability generalization: A systematic review and evaluation of meta-analytic methodology and reporting practice* (Unpublished dissertation, University of North Texas).
- *Holland, D. F., Kraha, A., Zientek, L. R., Nimon, K., Fulmore, J. A., Johnson, U. Y., Ponce, H. F., Aguilar, M. G., & Henson, R. K. (2018). Reliability generalization of the Motivated Strategies for Learning Questionnaire: A meta-analytic view of reliability estimates. *SAGE Open*, 8(3), 1–29. <https://doi.org/10.1177%2F2158244018802334>
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
- Jacoby, J., & Matell, M. S. (1971). Three-point Likert scales are good enough. *Journal of Marketing Research*, 8(4), 495–500. <https://doi.org/10.1177%2F002224377100800414>
- Kennedy, R. S., & Turnage, J. J. (1991). Reliability generalization: A viable key for establishing validity generalization. *Perceptual and motor skills*, 72(1), 297-298. <https://doi.org/10.2466%2Fpms.1991.72.1.297>


- Kettler, R. J., Rodriguez, M. C., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (2011). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. *Applied Measurement in Education, 24*(3), 210–234. <https://doi.org/10.1080/08957347.2011.580620>
- *Kilgus, S. P., Eklund, K., Maggin, D. M., Taylor, C. N., & Allen, A. N. (2018). The Student Risk Screening Scale: A reliability and validity generalization meta-analysis. *Journal of Emotional and Behavioral Disorders, 26*(3), 143–155. <https://doi.org/10.1177/1063426617710207>
- King, C., Phillips, C. E., Walker, K. D., & O'Toole, S. K. (2014). A reliability generalization of the attitudes towards women (AWS) scale. *Race, Gender, & Class, 21*(1–2), 151–168. <https://www.jstor.org/stable/43496966>
- Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care, 6*(1), 5–30. <https://doi.org/10.1017/S0266462300008916>
- *Lane, G., White, A., & Henson, R. (2002). Expanding reliability generalization methods with KR-21 estimates an RG study of the Coopersmith Self-Esteem Inventory. *Educational and Psychological Measurement, 62*(4), 685–711. <https://doi.org/10.1177/0013164402062004010>
- *Leach, L. F., Henson, R. K., Odom, L. R., & Cagle, L. S. (2006). A reliability generalization study of the Self-Description Questionnaire. *Educational and Psychological Measurement, 66*(2), 285–304. <https://doi.org/10.1177/0013164405284030>
- Leech, N. L., Onwuegbuzie, A. L., & O'Conner, R. (2011). Assessing internal consistency in counseling research. *Counseling Outcome Research and Evaluation, 2*(2), 115–125. doi:10.1177/2150137811414873
- *Leue, A., & Lange, S. (2011). Reliability generalization: An examination of the positive affect and negative affect schedule. *Assessment, 18*(4), 487–501. <https://doi.org/10.1177/1073191110374917>
- Lissitz, R. W., & Green, S. G. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60*(1), 10–13. <https://psycnet.apa.org/doi/10.1037/h0076268>
- Maeda, Y., & Rodriguez, M. C. (2002, April 1–5). *Statistical issues of reliability generalization and an application to achievement data* [Paper presentation]. Annual Meeting of the American Educational Research Association, New Orleans, LA, United States.
- *Mahapoonyanont, N., Krahamwong, R., Kochakornjarupong, D., & Rachasong, W. (2010). Critical thinking abilities assessment tools: Reliability generalization. *Procedia - Social and Behavioral Sciences, 2*(2), 434–438. <https://doi.org/10.1016/j.sbspro.2010.03.038>
- *Mji, A., & Alkhateeb, H. M. (2005). Combining reliability coefficients: Toward reliability generalization of the Conceptions of Mathematics Questionnaire. *Psychological reports, 96*(3), 627–634. <https://doi.org/10.2466/pr0.96.3.627-634>
- *Nilsson, J. E., Schmidt, C. K., & Meek, W. D. (2002). Reliability generalization: An examination of the Career Decision-Making Self-Efficacy Scale. *Educational and Psychological Measurement, 62*(4), 647–658. <https://doi.org/10.1177/0013164402062004007>
- Onwuegbuzie, A. J., & Daniel, L. G. (2000, November 17). *Reliability Generalization: The Importance of Considering Sample Specificity, Confident Intervals, and Subgroup Differences*. Paper presented at the Annual meeting of the Mid-South Educational Research Association. Bowling Green, KY.

- Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, *10*(2), 75-98. <https://doi.org/10.3102%2F10769986010002075>
- Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, *11*(3), 306.
- *Ross, M. E., Blackburn, M., & Forbes, S. (2005). Reliability generalization of the Patterns of Adaptive Learning Survey goal orientation scales. *Educational and Psychological Measurement*, *65*(3), 451-464. <https://doi.org/10.1177/0013164404272496>
- Ryngala, D. J., Shields, A. L., & Caruso, J. C. (2005). Reliability generalization of the Revised Children's Manifest Anxiety Scale. *Educational and Psychological Measurement*, *65*(2), 259-271. <https://doi.org/10.1177/0013164404272495>
- Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2012). 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.x>
- Semma, B., Henri, M., Luo, W., & Thompson, C. G. (2019). Reliability generalization of the meaning in life questionnaire subscales. *Journal of Psychoeducational Assessment*, *37*(7), 837-851. <https://doi.org/10.1177%2F0734282918800739>
- Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*(5), 529-540. <https://psycnet.apa.org/doi/10.1037/0021-9010.62.5.529>
- *Shou, Y. & Olney, J. (2020). Assessing a domain-specific risk-taking construct: A meta-analysis of reliability of the DOSPERT scale. *Judgment & Decision Making*, *15*(1), 112-134. <https://psycnet.apa.org/record/2020-10523-008>
- *Taylor, R. T. (2012). *Review of the Motivated Strategies for Learning Questionnaire (MSLQ) using reliability generalization techniques to assess scale reliability* [Unpublished doctoral dissertation]. Auburn University.
- Thompson, B. (2002). *Score Reliability: Contemporary Thinking on Reliability Issues*. Newbury Park, CA: Sage Publications.
- 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%2F0013164400602002>
- Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*(1), 6-20. <https://doi.org/10.1177/0013164498058001002>
- Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*(1), 562-569. <https://doi.org/10.1177%2F0013164402062004002>
- *Vacha-Haase, T., Kogan, L. R., Tani, C. R., & Woodall, R. A. (2001). Reliability generalization: Exploring variation of reliability coefficients of MMPI clinical scales scores. *Educational and Psychological Measurement*, *61*(1), 45-59. <https://doi.org/10.1177/00131640121971059>
- 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%2F00131640021970682>

- *Vacha-Haase, T., Tani, C. R., Kogan, L. R., Woodall, R. A., & Thompson, B. (2001). Reliability generalization: Exploring reliability variations on MMPI/MMPI-2 validity scale scores. *Assessment*, 8(4), 391–401. <https://doi.org/10.1177/107319110100800404>
- 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(3), 159–168. <https://doi.org/10.1177/0748175611409845>
- *Vassar, M., Knaup, K. G., Hale, W., & Hale, H. (2011). A meta-analysis of coefficient alpha for the Impact of Event Scales: A reliability generalization study. *South African Journal of Psychology*, 41(1), 6–16. <https://doi.org/10.1177/008124631104100102>
- Wang, J. (2002). Reliability generalization: An HLM approach. *Journal of Instructional Psychology*, 29(3), 213–219.
- *Warne, R. T. (2011). A reliability generalization of the Overexcitability Questionnaire–Two. *Journal of Advanced Academics*, 22(5), 671–692. <https://doi.org/10.1177/1932202X11424881>
- Wilder, L. K., & Sudweeks, R. R. (2003). Reliability of ratings across studies of the BASC. *Education and Treatment of Children*, 26(4), 382–399. <https://www.jstor.org/stable/42899768>
- Yetkiner, 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.
- 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. <https://doi.org/10.1177/00131640021970466>


Author Information

Ashley M. Williams

 <https://orcid.org/0000-0001-5131-1819>

Texas A&M University
801 Harrington Tower
College Station, TX 77843
United States

Jamaal Young

 <https://orcid.org/0000-0001-7277-1072>

Texas A&M University
801 Harrington Tower
College Station, TX 77843
United States

Contact e-mail: Jamaal.Young@tamu.edu
