SECTION FOUR

Results

Introduction

The purpose of this quasi-experimental study with between-group comparative design was to see if there is a connection between my SBG classroom (which uses differentiated reassessment) and higher metacognitive and motivational scores on the SMQII and PRO-SDLS surveys and whether this translated into larger statistical gains between SOCA and EOCA scores. Using the district's American History 1 SOCA and EOCA, the SMQII survey, and the PRO-SDLS survey, I ran statistical tests in the form of Principal Component Analysis (PCA), the Mann-Whitney U test, the Kruskal-Wallis H test, one-way ANOVA, and independent samples *t* tests in order to analyze the results and answer the three research questions. Data collection consisted of gathering student responses to the two surveys and their SOCA/EOCA scores from 13 American History classes at a suburban high school in Central Ohio.

Data Tracking

To track the data and results of both the SOCA/EOCA and the SMQII/PRO-SDLS, as well as to maintain confidentiality, I relied on the American History teachers in the other eight classes (five for Mr. E; three for Mr. O). Following the administration of the American History 1 and Honors American History 1 SOCA and EOCA, the individual Scantron forms were collected by the classroom teachers and scored, using the answer keys provided by the district. Since the EOCA was acting as a Final Exam for their courses, both teachers recorded the scores of each student into their online grade book. Only the SOCA and EOCA scores of students that had taken both forms of the assessment were used in this study. Those students that transferred classes or left the district and who did not take the EOCA had their SOCA scores removed.

Those students who arrived in the classroom after the SOCA had been given still took the EOCA for the purposes of their Final Exam grade, but their scores were also removed.

Principal Component Analysis: Do the Factor Structures of the SMQII and PRO-SDLS Conform to Expectations?

I used SPSS to run PCA independently for the SMQII and the PRO-SDLS. The hope in running PCA is to take a larger set of variables and condense them down into a smaller set of variables (the "principal components") that are strong enough to account for the discrepancy (or "variance") in the original data set. This allows the user to "cluster" variables together and reduce redundancy (Lund & Lund, 2016). Larger sample sizes are needed in order for PCA to yield reliable results, with most statistical sources recommending no fewer than 150 cases. As my study contained 232 survey participants and 327 test scores, sample size was not an issue. Additionally, PCA was conducted to make sure the resulting factors from my data matched up with Glynn's five factors in SMQII and Stockdale's four factors in PRO-SDLS.

Both the SMQII and the PRO-SDLS contain 25 questions, and both use a five-point scale. Knowing this, I elected to use a scale of 1 through 5 to code student responses. The SMQII used "never" (coded as 1), "rarely" (coded as 2), "sometimes" (coded as 3), "usually" (coded as 4), and "always" (coded as "5"). The PRO-SDLS used "strongly disagree" (coded as 1), "disagree" (coded as 2), "sometimes" (coded as 3), "agree" (coded as 4), and "strongly agree" (coded as 5). While this data is ordinal, Lovelace and Brickman (2013) caution against assuming that such coded numbers accurately represent units that can be transformed into data to justify the use of parametric statistical procedures. They caution that conclusions made in research that measure students attitudes, while impactful in studying educational strategies, "are only as good as the quality of the measures and the methods used to analyze the data collected" (p. 615). As a result, I needed to make sure that I produced support for the validity of my findings. Lovelace and Brickman further argue that individual questions must be treated as ordinal but become meaningful if the theoretically supported individual questions are broken into factors (composite scores for multiple questions). PCA would later show that my study's factor structure was valid, and because Glynn and Stockdale have theoretical support for their factors, I could justify the use of parametric statistics to draw conclusions from my students' responses.

Principal Component Analysis of SMQII and PRO-SDLS

Questionnaire data loaded into nine total factors. SMQII data loaded into five factors (F1-F5), while PRO-SDLS data loaded into four factors (F6-F9). These factors were: 1. *Intrinsic motivation* (F1); 2. *Career motivation* (F2); 3. *Self-determination* (F3); 4. *Self-efficacy* (F4); 5. *Grade motivation* (F5); 6. *Initiative* (F6); 7. *Control* (F7); 8. *Self-efficacy* (F8); and 9. *Motivation* (F9). Factors coded as F4 and F8 both included information pertaining to self-efficacy, but both were tied to different surveys (F4 with the SMQII; F8 with the PRO-SDLS). In terms of motivation, the SMQII divided this into three separate factors: intrinsic (F1), career (F2), and grade (F5). The PRO-SDLS did not distinguish different kinds of motivation and kept it within its own factor (F9).

The factor categories of the SMQII, along with the corresponding survey questions from my version (*American History Motivational Questionnaire* or AHMQ), can be found in Table 3.

Table 3

Principal Component Analysis: Factor loading of items from AHMQ

Factor 1 (F1). ***Intrinsic motivation***

Question 1. *The history I learn is relevant to my life.*

Question 3. *Learning history is interesting.*

Question 12. *Learning history makes my life more meaningful.*

Question 17. *I am curious about discoveries in history.*

Question 19. *I enjoy learning history.*

Factor 2 (F2). ***Career motivation***

Question 7. *Learning history will help me get a good job.*

Question 10. *Knowing history will give me a career advantage.*

Question 13. *Understanding history will benefit me in my career.*

Question 23. *My career will involve history.*

Question 25. *I will use history problem-solving skills in my career.*

Factor 3 (F3). ***Self-determination***

Question 5. *I put enough effort into learning history.*

Question 6. *I use strategies to learn history well.*

Question 11. *I spend a lot of time learning history.*

Question 16. *I prepare well for history tests.*

Question 22. *I study hard to learn history.*

Factor 4 (F4). ***Self-efficacy***

Question 9. *I am confident I will do well on history tests.*

Question 14. *I am confident I will do well on history assignments and projects.*

Question 15. *I believe I can master history knowledge and skills.*

Question 18. *I believe I can earn a grade of "A" in history.*

Question 21. *I am sure I can understand history.*

Factor 5 (F5). ***Grade motivation***

Question 2. *I like to do better than other students on history tests.*

Question 4. *Getting a good history grade is important to me.*

Question 8. *It is important that I get an "A" in history.*

Question 20. *I think about the grade I will get in history.*

Question 24. *Scoring high on history tests matters to me.*

Note. As cited in Glynn, Brickman, Armstrong, and Taaoobshirazi (2011, p. 1167).

The factor categories of the PRO-SDLS, along with the corresponding survey questions, can be found in Table 4:

Table 4

Principal Component Analysis: Factor loading of items from PRO-SDLS

Factor 6 (F6). ***Initiative*** Question 2. *I frequently do extra work in a course just because I am interested.* Question 9. *I would rather take the initiative to learn new things in a course rather than wait for the instructor to foster new learning.* Question 10. *I often use materials I've found on my own to help me in a course.* Question 15. *Even after a course is over, I continue to spend time learning about the topic.* Question 17. *I often collect additional information about interesting topics even after the course has ended.* Question 25. *I always rely on the teacher to tell me what I need to do in the course to succeed.*	**Factor 7 (F7).** ***Control*** Question 4. *If I am not doing as well as I would like in a course, I always independently make the changes necessary for improvement.* Question 5. *I always effectively take responsibility for my own learning.* Question 6. *I often have a problem motivating myself to learn.* Question 13. *I usually struggle in classes if the teacher allows me to set my own timetable for work completion.* Question 19. *I am very successful at prioritizing my learning goals.* Question 23. *I always effectively organize my study time.*
Factor 8 (F8). ***Self-efficacy***** Question 1. *I am confident in my ability to consistently motivate myself.* Question 7. *I am very confident in my ability to independently prioritize my learning goals.* Question 12. *I am very convinced I have the ability to take personal control of my learning.* Question 21. *I am really uncertain about my capacity to take primary responsibility for my learning.* Question 22. *I am unsure about my ability to independently find needed outside materials for my courses.* Question 24. *I don't have much confidence in my ability to independently carry out my student plans.* *** My survey data did not confirm Factor 8.*	**Factor 9 (F9).** ***Motivation*** Question 3. *I don't see any connection between the work I do for my courses and my personal goals and interests.* Question 8. *I complete most of my high school activities because I WANT to, not because I HAVE to.* Question 11. *For most of my classes, I really don't know why I take personal control of my learning.* Question 14. *Most of the work I do in my courses is personally enjoyable or seems relevant to my reasons for attending high school.* Question 16. *The primary reason I complete course requirements is to obtain the grade that is expected of me.* Question 18. *The main reason I do the course activities is to avoid feeling guilty or getting a bad grade.* Question 20. *Most of the activities I complete for my high school classes are NOT really personally useful or interesting.*

Note. As cited in Stockdale and Brockett (2011, p. 171).

However, before I could run PCA – and to guarantee that when I did I would be able to generate reliable results – I had to make sure that my variables were linearly related and that my data did not contain any outliers. Outliers can incorrectly skew the results, and SPSS Statistics recommends identifying outliers as scores greater than three standard deviations away from the mean (Lund & Lund, 2016; SPSS 23, 2016; Gravetter & Wallnau, 2008). My data and variables were linearly related and no outliers were found. In order to run PCA, my data had to also satisfy three tests: 1. The Kaiser-Meyer-Olkin (KMO) measure for the data sets within the SMQII and the PRO-SDLS surveys; 2. The KMO measure for each individual variable (25 variables for each survey); and 3. Bartlett's Test of Sphericity.

The KMO helps confirm whether or not a sample is sufficiently large to run a PCA on a data set by computing a value from 0 to 1, with the minimum requirement for sample size adequacy of 0.6. Values above 0.8 are considered "meritorious" on Kaiser's (1974) classification of measure values and indicate that a PCA will be useful. In my study, my version of the SMQII resulted in a KMO score of .900; my PRO-SDLS came in at .881. This suggests that my data sets for both surveys were sufficiently large for PCA. Table 5 and Table 6 display these results.

Table 5

KMO and Bartlett's Test for AHMQ

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.900
Bartlett's Test of Sphericity	Approx. Chi-Square	3441.681
	df	300
	Sig.	.000

Note. These tests were run using SPSS, version 23.

Table 6

KMO and Bartlett's Test for PRO-SDLS

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.881
Bartlett's Test of Sphericity	Approx. Chi-Square	2082.164
	df	300
	Sig.	.000

Note. These tests were run using SPSS, version 23.

Having met the KMO for my data sets, I then needed to look at the KMO value of my two surveys individual variables. With these scores, researchers would like for each of their variables to be as close to 1 as possible, with a minimum of 0.5. If an individual variable has a low KMO measure (KMO < 0.5), then it is recommended that researchers remove it from the analysis (Lund & Lund, 2016). In my study, all KMO measures came in greater than 0.7, meaning there was adequacy of sampling for the individual questions.

The final test to see if I was able to run PCA, called Bartlett's Test of Sphericity, checks to see if there are adequate correlations between variables. If there are adequate correlations between variables, it would mean my variables are capable of being reduced to a smaller number of components. (Lund & Lund, 2016). If Bartlett's Test of Sphericity was statistically significant ($p < .05$), then my data would be suitable for PCA. In referencing Table 5 on the previous page, this significance value (i.e., *p* value) is located in the "Sig." row. In both the SMQII and the PRO-SDLS, the Sig. shows ".000," which essentially means $p < .0005$. My data was found to be suitable for PCA.

The interpretability criterion for the Rotated Component Matrix shows how each component loads on each variable with an objective of achieving "simple structure." This means that each survey items loads strongly on only one factor, and each factor includes at least three survey items. In running these tests, my hope was that the 25 questions that I used from the SMQII and the 25 questions from the PRO-SDLS all load into the same pre-determined factors as loaded by Glynn (SMQII) and Stockdale (PRO-SDLS). Glynn's SMQII loaded Questions 1, 3, 12, 17, 19 into Factor 1 (*Intrinsic motivation*); Questions 7, 10, 13, 23, and 25 into Factor 2 (*Career motivation*); Questions 5, 6, 11, 16, and 22 into Factor 3 (*Self-determination*); Questions 9, 14, 15, 18, 21 into Factor 4 (*Self-efficacy*); and Questions 2, 4, 8, 20, and 24 into Factor 5 (*Grade motivation*). Stockdale's PRO-SDLS loaded Questions 2, 9, 10, 15, 17, and 25 into Factor 6 (*Initiative*); Questions 4, 5, 6, 13, 19, and 23 into Factor 7 (*Control*); Questions 1, 7, 12, 21, 22, and 24 into Factor 8 (*Self-efficacy*); and Questions 3, 8, 11, 14, 16, 18, and 20 into Factor 9 (*Motivation*).

The Rotated Component Matrix for the SMQII saw all of my questions for Factor 1 load in Factor 1, just like they did for Glynn. The same was true for Factors 2, 4, and 5. Question 11 (*"I spend a lot of time learning history"*) should have loaded for Factor 3 (*Self-determination*) but instead loaded for Factor 1 (*Intrinsic motivation*). This was the only question that did not match Glynn et al.'s results, so I used Glynn et al.'s structure. The Rotated Component Matrix for the PRO-SDLS saw Factors 6, 7, and 9 correctly load their corresponding questions. However, Factor 8 (*Self-efficacy*) had only three of its six questions land within the factor. Since Factor 8 from the PRO-SDLS was the same as Factor 4 of the SMQII, I elected to cut Factor 8 in favor of the results from Factor 4.

Are There Differences in Motivation and/or Metacognition in the Treatment and Non-Treatment Groups?

Motivation and Metacognition – All Students: SBG vs. Non-SBG

Having confirmed the SMQII and PRO-SDL factor structures, I was able to compare the coded student responses for Factors 1-5 (SMQII) and Factors 6-9 (PRO-SDLS) to see if there was a statistically significant difference between students in my classes (which utilizes SBG and differentiated reassessment) and those in Mr. E and Mr. O's classes (which do not utilize SBG and do not use differentiated reassessment). Since the independent samples *t* test is used when comparing two groups, I was unable to use that particular statistical test to examine three teachers. As such, I needed to use a one-way analysis of variance (ANOVA). However, according to Lund and Lund (2016), a one-way ANOVA is unable to tell a researcher which groups were statistically different – it can only show that at least two of them were different. With three groups in my study, determining which group is different is important, and follow-up post hoc tests were needed. Specifically, the Tukey post hoc provides statistical significance level (i.e., *p* value) for each comparison (Lund & Lund, 2016).

Using the Shapiro-Wilk Test for Normality, which confirms that a sample came from a normally distributed population (Lund & Lund, 2016), Factor 1 was found to be non-normal. This was confirmed because the skewness to standard error ratio for the data in Mr. E's class (also called the *z* score) was > 2.58. As a result, I did not worry about the two outliers present because this prompted me to use a nonparametric test. Factor 2 failed the Shapiro-Wilk Test, but I treated it as normal because all *z* scores were between –2.58 and 2.58. With Factor 3, only my students' scores failed the Shapiro-Wilk Test but passed the *z* score, which meant I could treat it as normal. The single outlier proved not to be an issue. Factors 4 and 5 were non-normal, but because I used a nonparametric test, the outliers were not an issue. When Factor 6 came back as normal, it had four outliers. The outliers were in nearly symmetric pairs in both my class and Mr. E's class, so I elected to keep the outliers. To make sure these four outliers did not impact the results, I re-ran the analysis twice (with and without the outliers), which confirmed that they did not. Factor 7 failed the Shapiro-Wilk Test but passed the *z* score tests for skewness and kurtosis, so I treated it as normal and included the outlier. Factor 8 failed the Shapiro-Wilk Test but passed the *z* score; however, Factor 8 was the factor I elected to cut because of the PCA and the inability of the questions to properly load on the factor. (Factor 4 replaced this factor.) Lastly, Factor 9 was normal but had two outliers. I ran the tests with and without the outliers to prove that they did not affect my results. They did not.

For the non-normally distributed data, I switched to the nonparametric Mann-Whitney U test for comparisons of two levels and the Kruskal-Wallis H test for comparison of three or more levels. With the normality tests, the one-way ANOVA was only valid for Factors 2, 3, 6, 7, and 9. For the others, Factors 1, 4, and 5, I ran the Kruskal-Wallis H test. For Factors 2, 3, 6, 7, and 9, equality of variance was satisfied for all variables, so I used the ANOVA table rather than the Robust Test for Equality of Means table. From the ANOVA table of my normally distributed variables, only Factor 3 had $p < .05$. This allowed me to conclude that there were no significant differences among classrooms for Factors 2, 6, 7, and 9. For Factor 3, however, there were some differences between classrooms, but because ANOVA is not capable of determining what those are, I turned to post-hoc tests, most notably the Tukey post hoc test. This test provides the researcher a statistical significance level (*p* value) for each pairwise comparison (Lund & Lund, 2016). Here I found that Mr. O (mean F3 = 3.07) and Mr. E (mean F3 = 3.538) were significantly different from one another (p = .133), and Mr. O (mean F3 = 3.071) and my class (mean F3 = 3.555) were significantly different from one another (p = .070). However, my class (mean F3 = 3.555) and Mr. E (mean F3 = 3.538) were not statistically different from one another. This did not support my hypothesis that my treatment causes a change, because if that were the case, then my class would have been significantly different from Mr. E's class. My class would also have been significantly different from Mr. O, while Mr. O and Mr. E would have been the same as one another. Not until I broke the classes into sub-groups and examined non-honors against non-honors and honors against honors did I find that Factor 5 (*Grade motivation*) was statistically significantly higher for my non-honors students (M = 4.384) than for the non-honors in Mr. E and Mr. O's non-SBG classrooms (M = 4.096), U = 1,318, p = .026. (This is discussed in greater detail in Section Five.)

As a result, I was unable to reject the first null hypothesis (H_{01}), meaning there is no statistically significant difference in the metacognitive effect of how 9th grade students (as a whole) perceive their learning in a social studies classroom that utilizes differentiated reassessment within a standards-based curriculum and students in traditional classrooms. I was also unable to reject the second null hypothesis (H_{02}), meaning there is no statistically significant difference in the motivational effect of how 9th grade students (as a whole) perceive their learning in a social studies classroom that utilizes differentiated reassessment within a standards-based curriculum and students in traditional classrooms. However, there is a statistically significant difference among non-honors students when it comes to *grade* motivation. Again, this is discussed at length in the next section.

Motivation and Metacognition – All Students: Non-Honors vs. Honors

Using all students that completed the SMQII and PRO-SDLS surveys (*N* = 232; 118 enrolled in non-honors classes and 114 enrolled in honors classes), the Shapiro-Wilk Test for Normality found Factors 3, 6, and 7 were normally distributed; Factors 1, 2, 4, 5, and 9 were not. A parametric independent samples *t* test indicated a statistically significant difference for Factor 3 (*Self-determination*) between honors (*M* = 3.660) and non-honors students (*M* = 3.295), *t* = -3.969, *p* = .0005. The test also found that for Factor 6 (*Initiative*), there were statistically significant differences between honors (*M* = 3.007) and non-honors students (*M* = 2.791); *t* = -2.918, *p* = .004). The tests found no statistically significant differences for Factor 7 (*Control*). The nonparametric Mann-Whitney U test, used for Factors 1, 2, 4, 5, and 9 found statistically significant differences between all honors and non-honors students for Factors 1, 4, and 5. Factor 1 (*Intrinsic motivation*) was statistically significantly higher for honors students (*M* = 3.546) than for the non-honors students (*M* = 3.102), *U* = 8,635, *p* = .0005. Factor 4 (*Self-efficacy*) was also statistically significantly higher for honors students (*M* = 4.067) than for the non-honors students (*M* = 3.647), *U* = 9,106, *p* = .0005. Lastly, Factor 5 (*Grade motivation*) was statistically significantly higher for honors students (*M* = 4.500) than for the non-honors students (*M* = 4.253), *U* = 8,653.50, *p* = .0005. The results can be seen in Table 7.

Table 7

Statistically Significant Factor Means for Honors vs. Non-Honors Students

Factor	Honors	Non-Honors	*p* values
Factor 1 (*Intrinsic motivation*)	M = 3.546	M = 3.102	$p = .0005$
Factor 3 (*Self-determination*)	M = 3.660	M = 3.295	$p = .0005$
Factor 4 (*Self-efficacy*)	M = 4.067	M = 3.647	$p = .0005$
Factor 5 (*Grade motivation*)	M = 4.500	M = 4.253	$p = .0005$
Factor 6 (*Initiative*)	M = 3.007	M = 2.791	$p = .0040$

Note 1. Factor 2 (*Career motivation*), Factor 7 (*Control*), and Factor 9 (*Motivation*) were not statistically significant. Factor 8 (*Self-efficacy*) was not used as my survey data did not confirm it.

Note 2. These tests were run using SPSS, version 23.

Motivation and Metacognition – Non-Honors Students: SBG vs. Non-SBG

Concerning the non-honors students that completed the SMQII and PRO-SDLS surveys (*N* = 118; 64 enrolled in my non-honors SBG classes and 54 in the non-honors non-SBG classes), the Shapiro-Wilk Test for Normality found Factors 1, 2, 3, 6 and 9 were normally distributed; Factors 4, 5, and 7 were not. My survey data did not confirm Factor 8 (*Self-efficacy*); as such, it was removed from analysis. (However, Factor 4, which also covered self-efficacy, was confirmed by my data, so it replaced Factor 8 in the analysis.) The parametric independent samples *t* test for Factors 1, 2, 3, 6, and 9 found that my SBG classroom does not affect non-honors students in a statistically significant way. The nonparametric Mann-Whitney U test, used for Factors 4, 5, and 7 because they were not normally distributed, was run to determine if there were differences in motivation and metacognition scores between non-honors students in SBG and non-SBG classrooms. Only Factor 5 (*Grade motivation*) was statistically significantly higher for my non-honors students ($M = 4.384$) than for the non-honors in the non-SBG classrooms ($M = 4.096$), $U = 1{,}318$, $p = .026$.

Motivation and Metacognition – Honors Students: SBG vs. Non-SBG

Concerning the honors students that completed the SMQII and PRO-SDLS surveys ($N = 114$; 53 enrolled in my honors SBG classes and 61 in the honors non-SBG classes), the Shapiro-Wilk Test for Normality found Factors 1, 2, 3, 6, 7 and 9 were normally distributed; Factors 4 and 5 were not. The parametric independent samples *t* test for Factors 1, 2, 3, 6, 7 and 9 and the nonparametric Mann-Whitney U test for Factors 4 and 5 both found that my SBG classroom does not affect honors students in a statistically significant way.

Are There Differences in SOCA/EOCA in the Treatment and Non-Treatment Groups?

The data analysis for the third question examined the difference (comparison) between two populations (those in my SBG classes; those in Mr. E and Mr. O's non-SBG classes). Sub-groups of students from each class were also analyzed, specifically how Mr. E's honors students performed when compared to my honors students and how Mr. O's and Mr. E's non-honors students scored when paralleled to my non-honors students. Does Mastery Teaching have an impact on the mastery of the material by 9th grade students?

As for the statistical tests, mean scores were computed and compared between the intervention and control groups using Hake Gains and independent samples *t* tests. Hake Gains look for "normalized gains" between pre- and post-test scores. Such a gain is the increase in each student's pre-test score divided by the average increase that would have resulted if each student had a perfect score on a post-test (Hake, 1998). Hake (1998) argues that the normalized gain is a meaningful measure of how well an intervention works when comparing populations having different pre-test scores. In my study, this intervention is SBG and differentiated reassessment. Such tests provided the right kind of statistical information to either reject or fail to reject the third null hypothesis (H_{03}).

The third research question for this study examined the difference in the EOCA scores of non-honors, honors, and all students combined within the two classrooms (SBG and non-SBG). Table 8 contains the SOCA and EOCA results by class, as well as the Hake Gains.

Table 8

Combined Group Statistics and Hake Gains for Non-SBG vs. SBG Classrooms

Group	*N*	SOCA*	EOCA	Hake	*t*	*p*
Non-SBG All Students	8	25.80	53.90	63.57%	1.679	.121
SBG All Students	5	28.68	58.22	71.49%		
Non-SBG Non-Honors	5	24.83	51.08	58.11%	3.125	.020
SBG Non-Honors	3	28.47	57.67	70.31%		
Non-SBG Honors	3	27.41	58.60	73.20%	–.002	.999
SBG Honors	2	29.00	59.05	74.40%		

* Both SOCA and EOCA scores were out of 70 possible points.

SOCA and EOCA – All Students: SBG vs. Non-SBG

After calculating the Hake Gains for all student EOCA scores, an independent samples *t* test was run to compare the Hake Gains percentages for the SBG and non-SBG classrooms. EOCA scores for each classroom (SBG vs. non-SBG) were normally distributed, as assessed by the Shapiro-Wilk Test ($p > .05$), and there was homogeneity of variances, as assessed by Levene's Test for Equality of Variances ($p = .042$). My SBG students did better on their EOCA (M = 71.49%, SD = .037) than the students enrolled in the non-SBG classes (M = 63.57%, SD = .094), a statistically significant difference, M = .075, 95% CI [-.02, .17], $t(11) = 1.679$, $p = .121$. The results for all student EOCA Hake Gain can be seen in Table 9.

Table 9

Independent Samples Test for All Students EOCA Hake Gain

	Levene's Test for Equality of Variances							95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower Bound	Upper Bound
Hake Gain	5.298	.042	1.679	11	.121	.0749538	.0446387	-.0232954	.1732029

Note. These tests were run using SPSS, version 23.

For all students, I was able to reject the third null hypothesis (H_{03}), meaning there is a statistically significant difference in the improvement between SOCA and EOCA scores of all students (combined) in a social studies classroom that utilizes differentiated reassessment within a standards-based curriculum and students in traditional classrooms.

SOCA and EOCA – Non-Honors Students: SBG vs. Non-SBG

After calculating the Hake Gains for non-honors student EOCA scores, an independent samples *t* test was run to compare the Hake Gains percentages for the non-honors classrooms. There were no outliers in the data, as assessed by inspection of a boxplot. EOCA scores for each non-honors classroom (SBG vs. non-SBG) were normally distributed, as assessed by the Shapiro-Wilk Test ($p > .05$), and there was homogeneity of variances, as assessed by Levene's Test ($p = .769$). My SBG non-honors students gained more between the pre- and post-test (M = 70.31%, $SD = .045$) than the non-honors students enrolled in the non-SBG classes (M = 58.11%, $SD = .056$), a statistically significant difference, $M = 0.12$, 95% CI [0.03, 0.21], $t(6) = 3.125$, $p = .020$, $d = .05$. The results for non-honors Hake Gain can be seen in Table 10.

Table 10

Independent Samples Test for Non-Honors EOCA Hake Gain

	Levene's Test for Equality of Variances							95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower Bound	Upper Bound
Hake Gain	.095	.769	3.125	6	.020	.1187120	.0379873	.0257604	.2116635

Note. These tests were run using SPSS, version 23.

For non-honors students, I was able to reject the third null hypothesis (H_{03}), meaning there is a statistically significant difference in the improvement between SOCA and EOCA scores of non-honors students in a social studies classroom that utilizes differentiated reassessment within a standards-based curriculum and non-honors students in traditional classrooms.

SOCA and EOCA – Honors Students: SBG vs. Non-SBG

After calculating the Hake Gains for honors student EOCA scores, an independent samples *t* test was run to compare the Hake Gains percentages for the honors classrooms. There were no outliers in the data, as assessed by inspection of a boxplot. EOCA scores for each honors classroom (SBG vs. non-SBG) were normally distributed, as assessed by the Shapiro-Wilk Test ($p > .05$), and there was homogeneity of variances, as assessed by Levene's Test ($p = .114$). My SBG honors students did not do statistically significantly better on their EOCA ($M = 74.4\%$, $SD = .015$) than the honors students enrolled in the non-SBG classes ($M = 73.2\%$, $SD = .061$), no statistically significant difference, $M = -.00008$, 95% CI [-0.15, 0.15], $t(3) = -.002$, $p = .999$. The results for honors EOCA Hake Gain can be seen in Table 11.

Table 11

Independent Samples Test for Honors EOCA Hake Gain

	Levene's Test for Equality of Variances							95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower Bound	Upper Bound
Hake Gain	4.907	.114	-.002	3	.999	-.0000814	.0460023	-.1464812	.1463184

Note. These tests were run using SPSS, version 23.

For honors students, I was unable to reject the third null hypothesis (H_{03}), meaning there is not a statistically significant difference in the improvement between SOCA and EOCA scores of honors students in a social studies classroom that utilizes differentiated reassessment within a standards-based curriculum and honors students in traditional classrooms.

Summary

I used 232 survey respondents to examine the first hypothesis (metacognition) and second hypothesis (motivation). One hundred and sixteen students were in my SBG classes while 116 were in the non-SBG classrooms (80 from Mr. E; 36 from Mr. O). This whole group comparison examined students in both honors and non-honors sections and cut across both SBG and non-SBG classes. I was unable to reject the first null hypothesis (H_{01}) and second null hypothesis (H_{02}), meaning there is no statistically significant difference in the metacognitive and motivational effect of how 9th grade students (as whole) perceive their learning in a social studies classroom that utilizes differentiated reassessment within a standards-based curriculum and students in traditional classrooms. However, there is a statistically significant difference among non-honors students when it comes to *grade* motivation.

All 327 student SOCA and EOCA scores were used to calculate averages for each American History 1 class's pre- and post-tests. Of the 327 students, 124 were in my SBG classroom and 203 were in the non-SBG classrooms (127 in Mr. E; 76 in Mr. O). Overall, my non-honors and honors students (combined) and my non-honors (individually) produced higher Hake Gains on their EOCA than the students in the non-SBG classes. I was able to reject the third null hypothesis (H_{03}), meaning there is a statistically significant difference in the improvement between SOCA and EOCA scores of students in a social studies classroom that utilizes differentiated reassessment within a standards-based curriculum and students in traditional classrooms.

Section Five advances the interpretation of my study's findings in greater detail. In addition, that section provides recommendations for action and presents my final reflection.