2004

Beyond Gender: A Meta-Analytic Approach to Gender Differences in Academic Achievement

Mary E. Nikityn

Southern Adventist University

Southern Scholars Honors Program

Beyond Gender:

A Meta-Analytic Approach to Gender Differences in Academic Achievement

Completed as a Southern Scholars Senior Project

by

Mary E. Nikityn

Faculty Advisor: Dr. Ruth WilliamsMorris

April 16, 2004

Contents

Beyond Gender:

A Meta-Analytic Approach to Differences in Academic Achievement

Mary E. Nikityn

Southern Adventist University

Abstract

Gender differences have long been researched in the arena of academic achievement, especially in math achievement. This study takes a meta-analytic approach to recent literature on gender differences in factors associated with mathematical performance. Relationships between such variables as gender, math achievement, self-concept, anxiety, stress, and interest were converted to Cohen's *d* for effect size and effect sizes are reported and compared. Current research demonstrates no conclusive trends in gender differences pertaining to any of these factors. Implications of findings are discussed and an agenda for further meta-analytic steps and future research is presented.

Beyond Gender:

A Meta-Analytic Approach to Differences in Academic Achievement

While it would probably be unfair to say that the modern domain of psychological research embraces certain "dead horse" areas of study, it is undeniably true that some research questions have been much more thoroughly investigated than others. Nor would it be difficult to make the case that the study of gender differences in mathematics performance is one of the most petrified of these farm animals. Hyde, Fennema, and Lamon, in a 1990 meta-analysis elegantly entitled "Gender Differences in Mathematics Performance" looked at no less than 100 studies on said topic. The results were still inconsistent. However, psychological perspectives aside, as a human being it is difficult to accept that one-half of the world's population is genetically condemned to perform poorly on math tasks. Surely it is necessary to look for other factors operating in the gender-math equation!

Rotational Ability, Anxiety, Gender, and Math Achievement

Casey et al. (1995, 1997) picked out mental rotational ability as a prominent one of these other factors. In two studies, researchers investigated the relationship between gender and the spatial ability to mentally rotate three-dimensional objects as a mediating factor in gender differing mathematical performance. Gender differences have also been found in anxiety (Ackerman, Bowen, Beier, & Kanfer, 2001; Casey, Nuttall, & Pezaris, 1997; Epstein, Pacini, Denes-Raj, & Heier, 1996) and anxiety differences have been found in relation to mathematical achievement (Viswanathan, 1993), although research has not yet emphasized a link between the two. Boekaerts (1996) did, however, present a framework wherein anxiety might be expected to adversely affect motivation and goal-directedness in learning.

Stress, Gender, and Academic Achievement

Not only anxiety, but also academic stress has been found to differ between genders. Epstein, Pacini, Denes-Raj, and Heier (1996) and Misra, Crist, and Burant (2003), in two different samples of undergraduate students, found significant gender differences in stress in college life and reaction to academic stressors, respectively. [In favor of which gender?]

Certainly it would be highly understandable for stress to be related to academic performance. Crocker, Karpinski, Quinn, and Chase (2003) researched a similar construct in a study of self-esteem and its relation to receiving good and bad grades.

Self-concept, Gender, and Achievement

Self-concept and self-esteem, then, also make up an important factor in understanding patterns of academic performance. Gender differences in self-concept—particularly in self-concept as it relates to perceived mathematical competence—have been found to varying degrees in a number of recent studies: Tiedemann (2000); Ackerman, Bowen, Beier, and Kanfer (2001); Pomerantz, Altermatt, and Saxon (2002); Vermeer, Boekaerts, and Seegars (2000); Bosacki (2000); and a 1999 meta-analysis by Kling, Hyde, Showers, and Buswell, which reported consistently higher self-esteem scores for boys than for girls. Martin, Marsh, and Debus (2001) linked lower self-esteem to defensive expectations of poorer academic performance.

Attitudes, Interests, and Achievement

Attitudes and interests must also play a role in academic performance—Viswanathan (1993) and Nosek, Banaji, and Greenwald (2002) studied aspects of attitudes toward math. Teacher attitudes toward student abilities were found to play a role in student self-concept by Jussim (1989). Webb, Lubinski, and Benbow (2002), Achter, Lubinski, and Benbow (1999), and Goff and Ackerman (1992) studied the role of interests in academic achievement. Bong (2001) elaborated further on this area of research emphasis by examining math task value and its role in achievement. Miller and Byrnes's (2001) study of the same year took a similar tack in examining the importance placed upon academic achievement in high school boys and girls. An interesting recent study by James and Richards (2003) even found differing patterns of academic interests in boys who attended single-sex and coed high schools.

Social Factors and Gender in Performance

The only quasi-experimentally designed study included in this research was undertaken by Inzlicht and Ben-Zeev in 2003: the performance of women on mathematical problems as minority group members was compared to the performance of women in entirely female groups. Similar studies found that group makeup had no effect on male performance, whereas minority females performed worse. This raises the question of whether gender differences exist in social understanding. Campbell and Williams (2000) researched differences in social consciousness and social desirability concerns, but clear gender differences were not found in their sample of ninth-grade students.

Additional Research on Cognitive Ability

Other lines of research relating to cognitive ability have been even more interesting and less conclusive: Davies, Stankov, and Roberts (1998) in three studies researched relationships between emotional intelligence, personality, and crystallized/social intelligence. Kuncel, Hezlett, and Ones (2004) studied the predictive ability of the Miller Analogies Test and were able to link general intelligence to both job and graduate school performance. Academic achievement in kindergarteners was the focus of a 2001 study by Kurdek and Sinclair: girls were found to be more skilled than boys in visuomotor skills, which predicted greater later mathematic achievement. Abele's 2003 study of gender differences in traits and social behavior found a lessening gap between masculine active, decisive traits and feminine caring, emotional traits.

Gender and Math Achievement

Naturally, though, despite the branching of research emphasis beyond simple gender studies, there has been no recent shortage in examinations of the original research question. Many studies, whatever their other variables, have continued to examine the relationship between gender and math performance. A number of these have been included as the backbone of this research: Casey, Nuttall, Pezaris, and Benbow (1995); Epstein, Pacini, Denes-Raj, and Heier (1996); Robinson, Abbott, Berninger, and Busse (1996); Casey, Nuttall, and Pezaris (1997); Swiatek, Lupkowski-Shoplik, & O'Donoghue (2000); Tiedemann (2000); Ackerman, Bowen, Beier, and Kanfer (2001); Pomerantz, Altermatt, and Saxon (2002); Webb, Lubinski, and Benbow (2002); and Penner (2003).

The purpose of this study was twofold: first, to examine the extent and consistency of gender differences in math performance in recent research and second, where gender differences are found, to look for mediating factors and variables other than gender (such as self-concept, anxiety, interest, etc) which might create such differences. This study was also intended to serve as a basis for future research on the role of non-gender aspects in mathematical achievement. It was hoped that such research would elucidate other factors that may contribute to differences in math ability and achievement.

Method

Materials

Thousands of research studies on some aspect of gender differences and math achievement exist. Of this wealth of available literature, 68 articles were identified as closely related to this research. Of these, 32 were believed to be especially relevant to the topic at hand and were selected for inclusion in this research (See list of studies in Appendix A). All articles were obtained using the PsychArticles database, available online through McKee Library at Southern Adventist University. The publication dates of included articles ranged from 1989 to 2004. Sample sizes ranged from 54 to 5,422.

Effect sizes were calculated by hand and using a scientific graphing calculator. Resulting data was graphed using Microsoft Office Excel.

Procedure

The researcher utilized the PsychArticles database to identify 68 recent research articles related to aspects of gender and math achievement. Upon in-depth reading, 32 of these were believed to be especially relevant to the topic at hand and were selected for inclusion in this research. These articles were thoroughly read and annotated, with their statistical procedures identified. Test results related to specific facets of gender and achievement were extracted and categorized into groups measuring the same relationship—i.e., gender and mathematics achievement, gender and self-concept, gender and anxiety.

Test statistics wherever applicable were converted to Cohen's *d* using a formula sheet (Lyons, n.d.). A number of studies already presented effect sizes, and one study presented Chi-square results, which could not be converted and thus have been reported as-is. Effect sizes have been presented and compared within study groups. For the gender and math achievement and gender and self-concept groupings, graphs have been used to further elucidate findings.

Statistical Analysis

All effect sizes have been presented using Cohen's *d*. A number of the studies used in this analysis reported effect sizes as Cohen's *d*. For studies without reported effect sizes, test statistics were converted to Cohen's *d* using meta-analytic formulas (Lyons, n.d.). Pearson's product moment correlations (*r*), one way analyses of variance (F), t-tests for independent samples (*t*), and means (X) were converted using separate formulas. Beta values for main effects of regression analysis (β) were treated as Pearson's *r* in calculation for the purposes of this analysis. Chi-square values (χ^2) have been reported as stated in the studies, due to the non-parametric nature of the statistical test.

The researcher elected to use Cohen's *d* for its fecundity in the comparison of a number of different statistical tests, as well as for its intrinsic distinction between small, medium, and large effect sizes. Thus, for the purposes of this study, an effect size of .20 will be considered a small effect size, .50 a medium effect size, and .80 a large effect size, in accordance with Cohen's stated interpretation of *d* for effect size (Aron & Aron, 2003).

Results

Gender and Math Achievement

A relatively recent meta-analysis of 100 studies on gender differences in mathematic performance stated that such differences are small and have declined over time since the 1970s (Hyde, Fennema, & Lamon, 1990). In 1996, Robinson, Abbott, Berninger, and Busse found a consistent pattern of gender differences in a high-performing preschool and kindergarten sample of 310, with more boys being nominated as high-performing, and boys in the sample performing better on mathematical evaluations than girls. The young age of this sample would seem to suggest an inherent, unsocialized gender difference. However, a range of recent studies has found a surprisingly inconsistent assortment of relationships between gender and math achievement. (See Figure 1.)

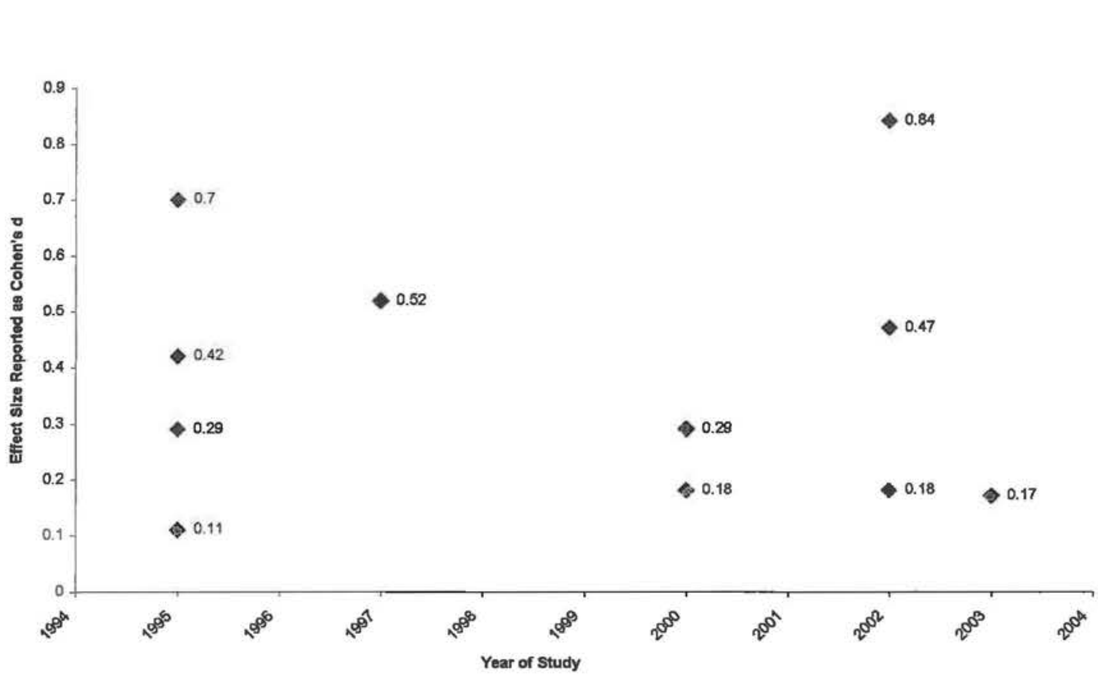

Figure 1. Effect sizes of gender on math achievement in seven studies, calculated as Cohen's *d*. Studies included are Casey, Nuttall, Pezaris, and Benbow (1995); Casey, Nuttall, and Pezaris (1997); Penner (2003); Pomerantz, Altermatt, and Saxon (2002); Swiatek, Lupkowski-Shoplik, and O'Donoghue (2000); Tiedemann (2000); and Webb, Lubinski, and Benbow (2002).

Penner, in 2003, used data from the 1995 Third International Mathematics and Science Study to study gender differences in math and science literacy in multiple countries. For the United States, a gender difference in math literacy of $d = .17$—a small effect size—was found, with males performing better than females. This effect size for the U.S. was the smallest gender difference of the ten countries in the study.

Additionally, Penner found that the gender difference increased with item difficulty, with the male advantage being greater on more difficult items than on easier items.

Swiatek and Lupkowski-Shoplik's 2000 sample of 5, 422 elementary school students in Pennsylvania and Pomerantz, Altermatt, and Saxon's 2002 sample of 932 elementary school students both found a similar small effect size for gender differences in academic performance: $d = .18$. Swiatek and Lupkowski-Shoplik found this difference on EXPLORE—a test for gifted elementary school students—scores in math, with boys performing better than girls, $F\,(1, 5412) = 34.50, p < .01$. Pomerantz, Altermatt, and Saxon examined gender differences in performance and internal distress, and found this effect size for gender differences in overall academic performance.

A range of effect sizes was found by Casey, Nuttall, Pezaris, and Benbow in a 1995 study. Separate sample groups of college students, talented preadolescents, and college-bound high school students each produced a different effect size for gender and math achievement as measured by SAT math scores. Among talented preadolescents, effect size was a medium to large $d = .70$; among high and low ability high school students, it was $d = .42$ and $d = .11$, respectively. College students themselves had a small to medium effect size of $d = .29$. This wide a range of effect sizes—from the very small to fairly large—throughout abilities and ages would seem to belie a consistent gender difference in math performance. However, when Casey, Nuttall, and Pezaris revisited the same arena of inquiry in 1997, a correlation of $r = .25, p < .05$ was found for gender differences on SAT math scores in 300 high school sophomores, with boys outperforming girls (boys' mean = 595, *SD* = 87; girls' mean = 554, *SD* = 81). This was a medium effect size of $d = .52$, although it is difficult to accept a single effect size for the entire sample following the great variation in subsample effect sizes in the 1995 study.

Among a younger student group—589 elementary school students in Germany—Tiedemann (2000) stated having found no significant gender differences in previous or current math grades. For previous math grades, $F\ (1, 27) = .57, p < .05$, which was a small effect size of $d = .29$. In current math grades, no difference whatsoever was found: $F\ (1, 27) = .00$. Nevertheless, as is discussed later, perceptions of students and teachers held boys to be more competent in math.

Much larger gender differences in math performance were found by Webb, Lubinski, and Benbow (2002) in a longitudinal sample of mathematically precocious youth. Participants were given the SAT at age 13 and follow up study was done at ages 18, 23, and 33. For both math-science participants and nonmath-nonscience participants (groups based on college major), men scored higher on the SAT math section than did women: $d\ (632, 258) = .84, p < .01$ and $d\ (126, 90) = .47, p < .01$, respectively. These were large and medium effect sizes, respectively. Although women in the sample had higher SAT verbal scores than men, due to the selective nature of the mathematically precocious sample, it would be unreasonable to generalize either of these areas of gender difference to a general population. The finding of a large effect size for gender differences in this group is, however, consistent with Casey et al's 1995 findings of greater gender differences in higher ability samples.

Among an even more highly specialized sample—university students who reported a need for cognition (as opposed to reliance on intuition)—Epstein, Pacini, Denes-Raj, and Heier (1996) found an interesting correlation. There was a significant relationship in women between cognitive orientation and performance on the SAT (math and verbal scales) of $r = .55$, $p < .001$. In men, the relationship between cognitive orientation and SAT performance was $r = .38$, $p < .001$. These were both large effect sizes of $d = 1.32$ for women and $d = .80$ for men. The authors reported a significant gender difference, although statistical tests supporting this were not presented.

Gender and Self-Concept

Studies have found gender differences in self-concept from very young ages (i.e., Bosacki, 2000). The question of how much self-concept (self-esteem, self-efficacy) can or does affect performance has yet to receive a definite answer. However, this of course has not stopped researchers from investigating and some inferences—albeit inconclusive—can perhaps be drawn. The range of effect sizes for gender and self-concept in recent studies has been wide, though not perhaps as wide as the range for gender and math achievement. (See Figure 2.) As with math, many of the differences in self-concept effect size must be addressed as pertinent to differing samples—both in type and quantity—and methods of measurement.

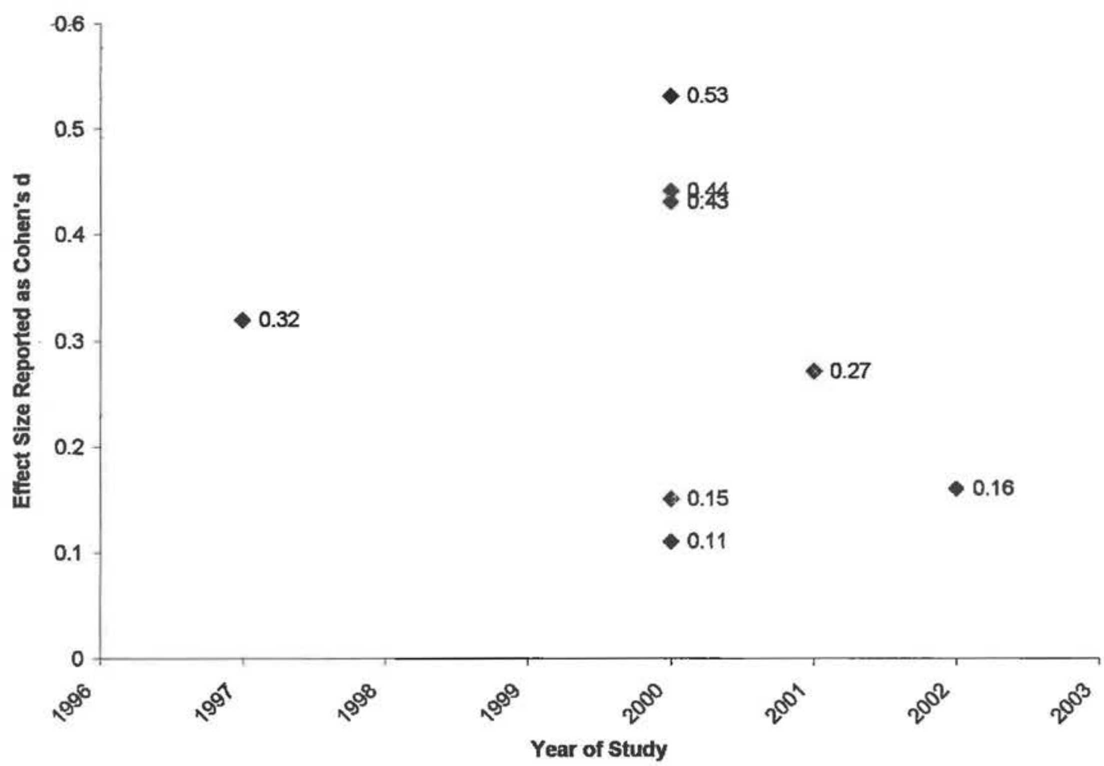

Figure 2. Effect sizes of gender on self-concept in six studies, calculated as Cohen's *d*. Studies included are Ackerman, Bowen, Beier, and Kanfer (2001); Bosacki (2000); Casey, Nuttall, and Pezaris (1997); Pomerantz, Altermatt, and Saxon (2002); Tiedemann (2000); and Vermeer, Boekaerts, and Seegars (2000).

For example, Bosacki (2001) found an effect size of only $d = .11$ for gender and global self-worth, with boys having slightly lower self-worth (M = 19.41, SD = 3.39) than girls (M = 19.81, SD = 3.72) in a sample of 128 11-year-olds. However, effect size for academic competence (academic self-concept) and gender was nearly a third larger, $d = .15$, with girls reporting lower levels than boys. The young age of this sample may account for its smaller effect sizes than other studies, as social roles may not yet have crystallized. Consistent with this interpretation are the results of a 1999 meta-analysis by Kling, Hyde, Showers, and Buswell: in measuring the magnitude of the gender difference in self-esteem in difference age groups, the researchers found the 7- to 10-year old age group to have an effect size of $d = .16$ for gender differences. In the 11- to 14-year old age group, this difference increased to $d = .23$, and in the 15- to 18-year old age group, to $d = .33$. (In later age groups, the effect size declined again, reaching a low of $d = -.03$ in the 60-plus age group.) This research highlights how effect size measurements of self-concept can vary greatly based on other, non-gender factors.

With that in mind, one can discuss gender differences in a number of recent studies related to self-concept. In a study of 589 elementary school children in Germany, Tiedemann (2000) found that boys perceived themselves to be more competent in mathematics than did girls, $F(1, 464) = 21.94, p < .001$. This is a medium effect size of $d = .43$. Interestingly, both parents and teachers also perceived boys as more competent in mathematics than girls, even though the study showed no gender differences in previous or actual math performance.

Vermeer, Boekaerts, and Seegars (2000) also found boys to report higher confidence levels than girls in their 158-student sixth-grade sample. Although there were no significant gender differences in computational mathematics problems, boys' confidence was significantly higher than girls' in two separate application problems: $t(156) = 3.34, p < .01$ and $t(156) = 2.27, p < .01$, respectively. These are effect sizes of $d = .53$ and $d = .44$, medium effect sizes. Also, although boys performed significantly better than girls on all application problems, girls were more likely to attribute bad results to their own lack of ability.

Pomerantz, Altermatt, and Saxon (2002), in their sample of 932 elementary school students, found an effect size of $d = .16$ for gender differences in self-evaluation. This was a small effect size, and very similar to Bosacki's 2001 finding of an effect size of $d = .15$ for gender differences in academic competence among 11-year-olds. Additionally, in the 2002 study, the effect size found for gender differences in self-evaluation was nearly the same as the effect size for gender differences in academic performance: $d = .18$. Both are small effect sizes, but their similar range is more striking than their size.

Casey, Nuttall, and Pezaris (1997) found a non-significant correlation of $r = .17$ between math self-confidence and gender in their sample of 300 high school sophomores taking the SAT. The effect size of this correlation was $d = .32$, a small to medium effect size. An effect size of similar small magnitude was found by Ackerman, Bowen, Beier, and Kanfer in their 2001 sample of 320 university freshmen: $d = -.27$. Men were found to have a higher math self-concept than women, $t\ (312) = -2.36, p < .05$. (In this case, a negative effect size denotes a higher mean score for men.)

Rotational Ability, Gender, and Math Achievement

Spatial reasoning has long been perceived as an important topic within the gender and math differences field of study. Do males have better spatial reasoning and rotational ability (the ability to mentally visualize the rotation of three-dimensional objects) than women? And would differences in this ability be consistent with whatever inconsistent differences in mathematics performance have been found? How much are math and rotational ability related? These were the questions asked by Casey, Nuttall, Pezaris, and Benbow in 1995, and revisited by Casey, Nuttall, and Pezaris in 1997.

In the 1995 study, college students, talented preadolescents, and college-bound high school students were organized into separate sample groups. The talented sample consisted of math precocious seventh- to ninth-grade students, and the high school group was divided into high- and low-ability samples based on SAT verbal scores. The effect size for gender difference in SAT math scores was medium to large, $d = .70$, for the talented preadolescent sample. The high-ability college bound sample had a medium effect size of $d = .42$, while the low-ability college bound sample had a much smaller effect size of $d = .11$. Gender differences in math performance in the college sample had a small effect size of $d = .29$. The study then looked at gender differences in mental rotational ability. In the talented preadolescent group, $d = .79$ for gender differences in mental rotation. The effect size was $d = 1.01$ for the college group and $d = .61$ for the high-ability sample. These are medium to quite large effect sizes. However, the gender difference in mental rotational ability for the low-ability college sample was only $d = .07$, a very, very small effect size.

When Casey et al statistically adjusted the gender differences in math performance to account for differences in mental rotation, the effect size was reduced to $d = .60$ for the talented sample, reduced to $d = -.01$ for the college sample, and reduced to $d = .16$ and $d = .08$ respectively for the high- and low-ability high school samples. Thus, mental rotational ability may account for at least some amount of gender differences in math performance.

In a later study, Casey, Nuttall, and Pezaris (1997) continued to find significant gender differences in both SAT math scores ($r = .25, p < .05; d = .52$) and in mental rotational ability ($r = .37, p < .01; d = .80$) with college bound high school boys

~~performing better than girls in both of these areas. It is interesting to note the persistence~~ of both findings and the much stronger effect size of gender difference in rotational ability than ~~in SAT math performance~~.

Gender, Anxiety, and Math Achievement

Gender differences are found in levels of anxiety as well as in self-concept and math achievement. Ackerman et al. (2001) found a slightly higher level of self-reported anxiety in women than in men among a sample of university freshmen: $t\,(314) = 1.67$, $p < .05$. This is a small effect size of $d = .18$. However, in a 1997 study, Casey, Nuttall, and Pezaris found a correlation of $r = .32$ ($p < .01$) between gender and math anxiety, with boys having significantly more math anxiety ($M = 29.10$, $SD = 8.06$) than girls ($M =$ 23.90, $SD = 7.34$). This is a medium effect size of $d = .68$.

Among women and men who reported a need for cognition (rather than a reliance on intuition), Epstein, Pacini, Denes-Raj, and Heier (1996) found a greater correlation between need for cognition and anxiety in men ($r = -.32$) than in women ($r = -.30$). Both correlations were significant at the $p < .001$ level and there was a significant gender difference between the two reported. For men, the effect size was $d = -.67$ while the effect size for cognition on anxiety in women was only $d = -.62$. Both of these are medium effect sizes.

A relationship has also been found between anxiety and math performance. Viswanathan (1993), in a study involving undergraduate university students, found significant correlations between achievement anxiety and grades in quantitative classes. Both a correlation of $r = .34$ between facilitating anxiety and grades and a correlation of $r = -.22$ and debilitating anxiety and grades were significant at the $p < .01$ level. This resulted in a calculated effect size of $d = .72$ for facilitating anxiety on grades—a medium to large effect size. An effect size of $d = -.45$ was calculated for debilitating anxiety on grades, which was a small to medium effect size. Although there has not been enough research published on the effect of anxiety on math grades to form any definite conclusions, Viswanathan's research demonstrates a calculable link between the two.

Thus, an effect of anxiety on grades is existent, but due to a lack of research emphasis on the differing effects of facilitating and debilitating anxiety its function remains unclear. Likewise, gender differences do exist in levels of academic and math anxiety. However, whether men or women are more affected by this anxiety is also unclear.

Gender and Stress

Although studies have not been done on aspects of stress and academic achievement, gender differences in levels of academic stress have been found in at least two recent studies. Misra, Crist, and Burant (2003) found that gender was related to reaction to stress in a sample of international students enrolled at U.S. universities. With greater amounts of academic stress, women had higher reactions to stressors: $\beta = .27$. With β treated as r in calculation, $d = .56$, a medium effect size. Epstein et al (1996) found that among sampled university students who reported a need for cognition, there were significant gender differences in levels of stress in college life: $r = -.33$ for men and $r = -.13$ for women. For men, this correlation is significant at the $p < .001$ level, and for women it is significant at the $p < .05$ level. The effect size is relatively large, $d = -.70$, for men and small, $d = -.26$, for women. Although the body of available research cannot yield any conclusions about the relationship between academic stress and mathematical performance or what role gender differences might play, strong (though inconsistent) gender differences have been found in academic stress among students.

Stereotyping, Gender, Achievement, and Self-Concept

Stereotyping and the role of self-fulfilling prophecy on math performance have also been examined from a number of different perspectives. Jussim (1998) found teachers' perceptions of students' ability to have a direct effect on student self-concept of math ability, $\beta = .11, p < .05$. With β treated as r for calculation, the effect size of this relationship was $d = .22$, which is a small effect size. Jussim in turn found student self-concept of ability to have a direct effect on student grades, $\beta = .24, p < .05$. The effect size of this, again with β treated as r, was calculated as $d = .49$—a medium effect size. Student gender was found to have a slight relationship to teachers' perceptions of talent and effort, with boys perceived as more talented ($\beta = .07, p < .05$) and girls perceived as putting forth more effort ($\beta = -.15, p < .05$). These are effect sizes of $d = .14$ and $d = -.30$, respectively. Both are small effect sizes, and gender was not found to have a relationship with student grades.

Forms of self-stereotyping are also related to aspects of mathematical performance. Bong (2001), in a study of 424 Korean middle and high school students, found in the high school sample a correlation of $r = .74$, significant at the $p < .05$ level, between the value students placed on math and their math self-efficacy. A similar correlation was found in the middle school sample: $r = .71, p < .05$. The effect size of this relationship was $d = 2.20$ for the high school sample and $d = 2.02$ for the middle school sample, which are very large effect sizes. Although the high significance of these correlations may be called into question by the number of variables (20) present in the study's correlation matrix, it is certainly plausible that there would be a strong relationship between the value placed on math and a student's self-perceived ability in it.

Although Bong did not address gender differences, a study by Nosek, Banaji, and Greenwald (2002) found an interesting relationship. Among male and female undergraduate students in math-intensive majors, female students had more negative attitudes toward math than men: $t\,(67) = -2.97, p = .004$. For gender and negativity toward math within math-intensive majors, there was an effect size of $d = .73$, which is a large effect size. Thus, even for women highly involved in mathematical study, there was a significantly lower value placed upon mathematics than for men studying math.

Defensive expectations in math were also examined in a 2001 study by Martin, Marsh, and Debus. A correlation of $r = -.17, p < .05$ was found between defensive expectations in math and self-esteem in a sample of 584 Australian teacher education students. This was a medium effect size of $d = -.45$. Thus, self-esteem is lower for defensive pessimists. Although gender differences in defensive pessimism were not examined, gender differences in self-esteem related to academic grades were studied by Crocker, Karpinski, Quinn, and Chase (2003). Baseline gender differences in self-esteem were found to be $t\,(103) = .87, p = .383$. The calculated effect size for this difference was $d = .17$, a small effect size. The study also looked at gender differences in the relationship between good and bad grades and self-esteem, but these differences were even smaller than baseline differences.

A 2003 quasi-experimental study by Inzlicht and Ben-Zeev found that when given a math test to complete, female undergraduate students who were in groups of three with two male confederates performed worse than women who were in groups of three with two other women: F (1, 49) =6.97, p < .02. The effect size for this was d = .70, a medium effect size. The study conjectured that females were threatened by being minorities in a negatively stereotyped environment (i.e., the assumption that women perform inferiorly in math), and that—since altering the level of privacy of the test results did not affect differences in performance—females may have internalized these negative stereotypes.

Interest, Gender, Achievement, and Type of Education

Achter, Lubinski, Benbow, and Eftekhari-Sanjani (1999), in analyzing data from participants in a longitudinal study at ages 13 and 23, found gender differences not to be a factor in math achievement, and recommended the encouragement of individual interests. While this certainly seems to clear things up considerably, also raises the question of whether there are gender differences in interests. James and Richards (2003) surveyed male alumni from 12 different U. S. high schools and arrived at the surprising finding that boys who graduated from single-sex high schools chose humanities majors in college significantly more often than did boys who had graduated from coed high schools: χ^2 (2, N = 412) = 10.62, p < .01. Although this does not address the issue of whether female students are directed away from pursuing mathematics interests, it does present evidence of gender-specific discipline-directing in coed high schools. (Interestingly, James and Richards (2003) found no similar differences in math/science or business major groups.)

Thus, there is a possibility that girls are subtly directed away from pursuit of math interests. However, there also remains the possibility that female students simply have less interest in math-base disciplines than do male students. Nosek, Banaji, and

Greenwald (2002) found that women in their sample of Yale undergraduates in a psychology class had, regardless of their major, more negative attitudes toward math than did men, regardless of their major: $\beta = -.33$, $p < .0001$. In fact, there was a stronger dislike for math among women versus men than there was among non-math majors versus math majors ($\beta = -.17$, $p < .005$). This was a medium to large effect size of $d = -.70$.

In another telling study, Webb, Lubinski, and Benbow (2002)—using the same longitudinal sample as Achter et. al (1999)—found that among math and science majors, men took significantly more math and science classes than did women: d (632, 258) = .64, $p < .01$. This points to the inference that perhaps even women who are quantitatively talented and pursuing careers in math fields exhibit less interest in it than do men of similar educational vocation. In their discussion, Webb, Lubinski, and Benbow addressed their findings by citing a the possibility that mathematically talented women are more verbally talented than men who are equally mathematically talented, and choose to pursue fields that utilize their verbal skills. (As an interesting aside, this study also found that a greater percentage of women, 7.6, who had math-science degrees listed "homemaker" as their occupation than women who obtained nonmath-nonscience degrees—only 3.9 percent.)

Unfortunately for women who are not interested in math and science, relationship between these interests and math achievement have been found. Viswanathan (1993) found a correlation of $r = .24$, $p < .01$ between attitude toward mathematics and average grades in quantitative courses in a sample of midwestern university students. This was an effect size of $d = .49$, a medium effect size. Goff and Ackerman (1992) also found a

r^2: 0.0576 ?

correlation between ACT math scores and interests in their sample of undergraduate students. There was a correlation of $r = .22$ between interest in science and ACT math score, and a correlation of $r = .26$ between interest in technology and ACT math. Interest in math was not a variable. These are effect sizes of $d = .45$ and $d = .53$, respectively; both are medium effect sizes.

However, another existent gender difference is often overlooked as it pertains to math achievement. Interest in achievement itself varies between genders. Miller and Byrnes (2001) found that academic achievement was more important to ninth-grade students than to eleventh-grade students, with a medium effect size of $d = .47$. Additionally, academic achievement among eleventh-grade students was rated as much more important by girls than by boys, with a medium to large effect size of $d = .72$. What relationship these different levels of achievement foci may have to student math performance is uncertain, but any level of non-ability gender difference is elucidating to the study at hand.

Discussion

What, then, is the role of gender in mathematical performance, and what are the roles of the other factors that have been examined? At this stage of research, it is difficult if not impossible to ascertain clear answers. Although this study has taken a meta-analytic approach to the topic at hand, certainly the importance of the topic demands a full follow-through of all the meta-analytic steps. The nature, level, and time constraints of this project prevented completion of a meta-analysis in entirety. However, any agenda for future research should certainly include the final stage of meta-analytic research: further coding of the included studies for research design factors and characteristics of sample populations, a complete reliability check of all coded data, grouping of independent and dependent variables, calculation of mean and variance of effect sizes across studies, searching for moderator variables using Chi-square significance testing, and determining the mean and variance of effect sizes within moderator subgroups. Due to the broad range of sample populations, measures of achievement, and test statistics used in the studies included here, such follow-up work would be invaluable in crystallizing and clarifying the information gained in this preliminary analysis.

Certainly, too, the role of further research in elucidating the factors involved in math achievement cannot be overemphasized. In the artificiality of simple gender divisions, intragender variability has been considerably overlooked. The relationships between age, talent, and levels of gender difference have been one of the few points of consistency within previous research. With this in mind, it would seem only logical to call for further research that attempts to understand these factors (age and talent) and their role in math achievement.

Personality, too, is a non-gender variable that would be of benefit included in future math achievement studies. Although studies have often—especially in the early 1990s—examined the role of personality in job success, and at least a few studies have linked personality traits to intelligence measures, little to no knowledge base exists on what role personality type might play in mathematical or even simply academic achievement. With Kuncel, Hezlett, and Ones's 2004 study linking both job performance and graduate school performance to general intelligence on the Miller Analogies test, these areas of research are becoming increasingly interconnected—a trend which cannot help but be of benefit to psychology's ongoing study of academic achievement.

However, whatever the causes of gender differences in mathematical and general academic achievement, the very existence of these differences, even at highly inconsistent levels, is cause for concern. Although Ackerman, Bowen, Beier, and Kanfer (2001) found a distinct math knowledge advantage for males over females, this was not their most disturbing result. In fact, knowledge differences favoring men were found in nearly all areas of academic study, even those areas in which females stereotypically outperform males. Ironically enough, psychology knowledge was the only variable in which women had a very slight, insignificant, but distinctly present advantage.

Another prong of this same study (Ackerman, Bowen, Beier, & Kanfer, 2001) used data from the May 2000 administration of Advanced Placement (AP) tests to high school students. Although 574,905 tests in a variety of academic disciplines were administered to males and 667,419 tests were administered to females, less females than males obtained clear passing scores. Clear passing scores were obtained by 225,575 men, but only 217,572 women—even though 92,514 more women took the tests, many of which were in areas such as English literature that stereotypically show female advantages. This fact, that girls demonstrate worse performance in these academic areas where they should (according to other research) perform better than males, raises the issue of what role socialization might play in female academic achievement and test performance. What aspects of the high-stakes environment of AP tests might contribute to worse female performance, and what are the repercussions and implications for female students who are consequently required to take more college coursework than their male counterpart? If female students perform worse than males in areas in which they should perform better, perhaps gender differences in academic achievement are indeed less of a factor than gender differences in other areas of functioning.

However, despite the existence of some level of gender differences in many areas of psychological inquiry, it is difficult to excuse the inherent artificiality of gender comparisons. Although gender makes a simple and convenient independent variable in a wide range of research designs, it is important to remember that there will always be more variation within a gender than between the two genders. Factors such as age, ability, personality, and, increasingly, sexual orientation, interact in ways that defy dichotomous gender differentiation. At the end, perhaps only two conclusions can be safely reached: the interconnected variables related to human performance are too complex to be safely analyzed as simply a product of gender, and, as always, further research is needed.

References

Abele, A. E. (2003). The dynamics of masculine-agentic and feminine-communal traits: Findings from a prospective study. *Journal of Personality and Social Psychology, 85*, 768-776. Retrieved January 29, 2004 from PsychArticles.

Achter, J. A., Lubinski, D., Benbow, C. P., & Eftekhari-Sanjani, H. (1999). Assessing vocational preferences among gifted adolescents adds incremental validity to abilities: A discriminant analysis of educational outcomes over a 10-year interval. *Journal of Educational Psychology, 91*, 777-786. Retrieved January 29, 2004 from PsychArticles.

Ackerman, P. L., Bowen, K. R., Beier, M. E., & Kanfer, R. (2001). Determinants of individual differences and gender differences in knowledge. *Journal of Educational Psychology, 93*, 797-825. Retrieved January 15, 2004 from PsychArticles.

Aron, A., & Aron, E. N. (2003). *Statistics for Psychology* (3rd. ed.). Upper Saddle River, New Jersey: Pearson Education, Inc.

Boekaerts, M. (1996). Self-regulated learning at the juncture of cognition and motivation. *European Psychologist, 1*, 100-112. Retrieved January 29, 2004 from PsychArticles.

Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task-value, and achievement goals. *Journal of Educational Psychology, 93*, 23-24. Retrieved January 29, 2004 from PsychArticles.

Bosacki, S. L. (2000). Theory of mind and self-concept in preadolescents: Links with gender and language. *Journal of Educational Psychology, 92*, 709-717. Retrieved January 29, 2004 from PsychArticles.

Campbell, K. J., & Williams, J. P. (2000). Readers' social desirability and text that violates social values: Evidence of an interaction. *Journal of Educational Psychology, 92*, 515-523. Retrieved January 29, 2004 from PsychArticles.

Casey, M. B., Nuttall, R. L., & Pezaris, E. (1997). Mediators of gender differences in mathematics college entrance test scores: A comparison of spatial skills with internalized beliefs and anxieties. *Developmental Psychology, 33*, 669-680. Retrieved January 13, 2004 from PsychArticles.

Casey, M. B., Nuttall, R., Pezaris, E., & Benbow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology, 31*, 697-705. Retrieved January 29, 2004 from PsychArticles.

Crocker, J., Karpinski, A., Quinn, D. M., & Chase, S. K. (2003). When grades determine self-worth: Consequences of contingent self-worth for male and female engineering and psychology majors. *Journal of Personality and Social Psychology, 85*, 507-516. Retrieved January 29, 2004 from PsychArticles.

Davies, M., Stankov, L., & Roberts, R. D. (1998). Emotional intelligence: In search of an elusive construct. *Journal of Educational Psychology, 75*, 989-1015. Retrieved January 29, 2004 from PsychArticles.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of*

Personality and Social Psychology, 71, 390-405. Retrieved January 13, 2004 from PsychArticles.

Goff, M., & Ackerman, P. L. (1992). Personality-intelligence relations: Assessment of typical intellectual engagement. *Journal of Educational Psychology, 84*, 537-552. Retrieved January 15, 2004 from PsychArticles.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139-155. Retrieved January 13, 2004 from PsychArticles.

Inzlicht, M., & Ben-Zeev, T. (2003). Do high-achieving female students underperform in private? The implications of threatening environments on intellectual processing. *Journal of Educational Psychology, 95*, 796-805. Retrieved January 13, 2004 from PsychArticles.

James, A. N., & Richards, H. C. (2003). Escaping stereotypes: Educational attitudes of male alumni of single-sex and coed schools. *Psychology of Men and Masculinity, 4*, 136-148. Retrieved February 23, 2004 from PsychArticles.

Jussim, L. (1989). Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology, 57*, 469-480. Retrieved January 13, 2004 from PsychArticles.

Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin, 125*, 470-500. Retrieved February 23, 2004 from PsychArticles.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all?

Journal of Personality and Social Psychology, 86, 148-161. Retrieved January 29, 2004 from PsychArticles.

Kurdek, L. A., & Sinclair, R. J. (2001). Predicting reading and mathematics achievement in fourth-grade children from kindergarten readiness scores. *Journal of Educational Psychology, 93*, 451-455. Retrieved January 29, 2004 from PsychArticles.

Lyons, L. C. (n. d.). *Meta-analysis: Methods of accumulating results across research domains*. Retrieved April 14, 2004 from http://www.lyonsmorris.com/MetaA/macalc/MApaper.pdf

Martin, A. J., Marsh, H. W., & Debus, R. L. (2001). Self-handicapping and defensive pessimism: Exploring a model of predictors and outcomes from a self-protection perspective. *Journal of Educational Psychology, 93*, 87-102. Retrieved January 29, 2004 from PsychArticles.

Miller, D. C., & Byrnes, J. P. (2001). To achieve or not to achieve: A self-regulation perspective on adolescents' academic decision making. *Journal of Educational Psychology, 93*, 677-685. Retrieved January 29, 2004 from PsychArticles.

Misra, R., Crist, M., & Burant, C. J. (2003). Relationships among life stress, social support, academic stressors, and reactions to stressors of international students in the United States. *International Journal of Stress Management, 10*, 137-157. Retrieved January 15, 2004 from PsychArticles.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology, 83*, 44-59. Retrieved January 29, 2004 from PsychArticles.

Penner, A. M. (2003). International gender x item difficulty interactions in mathematic and scientific achievement tests. *Journal of Educational Psychology, 95*, 650-655. Retrieved January 29, 2004 from PsychArticles.

Pomerantz, E. M., Altermatt, E. Y., & Saxon, J. L. (2002). Making the grade but feeling distressed: Gender differences in academic performance and internal distress. *Journal of Educational Psychology, 94*, 396-404. Retrieved January 15, 2004 from PsychArticles.

Robinson, N. M., Abbott, R. D., Berninger, V. W., & Busse, J. (1996). The structure of abilities in math-precocious young children: Gender similarities and differences. *Journal of Educational Psychology, 88*, 341-352. Retrieved January 29, 2004 from PsychArticles.

Swiatek, M. A., Lupkowski-Shoplik, A., & O'Donoghue, C. C. (2000). Gender differences in above-level EXPLORE scores of gifted third through sixth graders. *Journal of Educational Psychology, 92*, 718-723. Retrieved January 29, 2004 from PsychArticles.

Tiedemann, J. (2000). Parents' gender stereotypes and teachers' beliefs as predictors of children's concept of their mathematical ability in elementary school. *Journal of Educational Psychology, 92*, 144-151. Retrieved January 29, 2004 from PsychArticles.

Vermeer, H. J., Boekaerts, M., & Seegers, G. (2000). Motivational and gender differences: Sixth-grade students' mathematical problem-solving behavior. *Journal of Educational Psychology, 92*, 308-315. Retrieved January 29, 2004 from PsychArticles.

Viswanathan, M. (1993). Measurement of individual differences in preference for numerical information. *Journal of Applied Psychology, 78*, 741-752. Retrieved January 13, 2004 from PsychArticles.

Webb, R. M., Lubinski, D., & Benbow, C. P. (2002). Mathematically facile adolescents with math-science aspirations: New perspectives on their educational and vocational development. *Journal of Educational Psychology, 94*, 785-794. Retrieved January 13, 2004 from PsychArticles.

Author's Note

Mary E. Nikityn, School of Education and Psychology, Southern Adventist University. This work has been completed as a Southern Scholars Honors Program Senior Project.

The author would like to thank Dr. Ruth WilliamsMorris for her consistent help and support throughout the course of this research endeavor. The researcher would also like to thank Southern Adventist University's Information Systems for failing to set up the pay printing system in the Hickman Science Center computer lab, thusly allowing her to print out several hundred pages of research articles for free. The researcher is sorry to note that said pay printing system was installed before she could print out this paper in quadruplicate.

Correspondence concerning this study should be addressed to Mary E. Nikityn, 6 Holly Road, Columbia, NJ 07832. Email: mnikityn@southern.edu.

Appendix A

Table 1.
Studies Used and Their Populations in Chronological Order

Year Published	Author(s)	*N* Size	*N* Type
1989	Jussim	27, 429	Teachers and their 6th grade students, respectively
1990	Hyde, Fennema, & Lamon	Meta-analysis	
1992	Goff & Ackerman	147	University of Minnesota undergraduates
1993	Viswanathan	93, 160, 90	Midwestern university students in 3 studies
1995	Casey, Nuttall, Pezaris, & Benbow	760	High- and low-ability college bound students taking SAT
1996	Boekaerts	Conceptual theory	
1996	Epstein, Pacini, Denes-Raj, & Heier	184, 973	Undergraduate psychology students in 2 studies
1996	Robinson, Abbott, Berninger, & Busse	310	High math-scoring preschoolers and kindergarteners
1997	Casey, Nuttall, & Pezaris	300	High school sophomores taking the SAT
1998	Davies, Stankov, & Roberts	100. 300, 131	Psychology students and Air Force recruits in 3 studies
1999	Achter, Lubinski, & Benbow	1110	Longitudinal study participants at ages 13 and 23
1999	Kling, Hyde, Showers, & Buswell	Meta-analysis	
2000	Bosacki	128	Male and female preadolescents, mean age 11
2000	Campbell & Williams	72	Ninth graders, 23 boys and 49 girls
2000	Swiatek, Lupkowski-Shoplik, & O'Donoghue	5422	Third through sixth-graders in Pennsylvania
2000	Tiedemann	589	Elementary school students in Germany
2000	Vermeer, Boekaerts, & Seegars	158	Sixth grade students
2001	Kurdek & Sinclair	281	Longitudinal study participants, kindergarten and grade 4
2001	Ackerman, Bowen, Beier, & Kanfer	320, Archival	University freshmen, archival AP test scores
2001	Miller & Byrnes	421	High school boys
2001	Bong	424	Korean middle and high-school students
2001	Martin, Marsh, & Debus	584	Austrialian teacher education students, mean age 20
2002	Nosek, Banaji, & Greenwald	83, 97	Two studies using Yale undergraduates
2002	Pomerantz, Altermatt, & Saxon	932	Elementary school students
2002	Webb, Lubinski, & Benbow	1110	Same longitudinal participants as Achter et al. (1999)
2003	Penner	Archival	Data from Third International Math & Science Survey
2003	Inzlicht & Ben-Zeev	54	High-math achieving female university students
2003	Crocker, Karpinski, Quinn, & Chase	122	Male and female engineering and psychology majors
2003	Misra, Crist, & Burant	143	International university students in US
2003	James & Richards	412	Male alumni from 12 US high schools
2003	Abele	1868	German university students
2004	Kuncel, Hezlett, & Ones	Meta-analysis	

Appendix B

$$\frac{\sum N r_i}{N_i}$$

Meta - Analysis: Methods of Accumulating Results Across Research Domains

Larry C. Lyons
Manassas VA
email solomon@mnsinc.com
home page: www.mnsinc.com/solomon/MetaAnalysis.html
Larry C. Lyons is currently with the George Washington University Medical Center.
Ross 312, 2300 I St. NW
Washington, DC 20037
tel: 202.994.2264
fax: 202.994.3513

e-mail: bmclxl@gwumc.edu

Abstract

This paper describes the Hunter-Schmidt method of conducting a Meta-Analysis. Meta-analysis is a set of statistical procedures designed to accumulate experimental and correlational results across independent studies that address a related set of research questions. The paper gives a brief description of meta-analysis methods based on procedures suggested by Hunter, Schmidt, and Jackson (1982) and Hunter and Schmidt (1990). It also presents the formulas and procedures needed for converting study statistics to a common metric, calculating the sample weighted mean $\underline{r}$ and $\underline{d}$, and correcting for range restriction and sampling and measurement error.

Meta - Analysis: Methods of Accumulating Results Across Research Domains

Introduction

Meta-analysis is a set of statistical procedures designed to accumulate experimental and correlational results across independent studies that address a related set of research questions. Unlike traditional research methods, meta-analysis uses the summary statistics from individual studies as the data points. A key assumption of this analysis is that each study provides a differing estimate of the underlying relationship within the population (rho). By accumulating results across studies, one can gain a more accurate representation of the population relationship than is provided by the individual study estimators.

Glass and colleagues (e.g., Glass, 1976; 1977; Glass &;Smith, 1977; McGaw &;Glass, 1980; Smith &;Glass, 1977; and Smith, Glass &;Miller, 1980) coined the term meta-analysis, and introduced most of the currently used procedures to psychology.

> *Meta-analysis refers to the analysis of analyses . . . the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature.*
>
> (Glass, 1976, p 3).

There are two general types of Quantitative Review procedures. One method involves the combination of probability values or Z scores, while the second technique combines effect sizes, such as Cohen's d (Cohen 1977, 1988) and the correlation coefficient, r. The procedures for combining Z or probability values was developed in parallel during the 30's by Cochran (1937), Fisher (1932), Pearson (1933) and Tippett (1931). These procedures were developed to address the need in agricultural research to combine the results of a number of independent tests, all of which were planned to test a common hypothesis. An alternative approach was also developed by Fisher in 1932, the r to Z transformation.

The demands of World War II served to assist in the development of combinatorial procedures. In their landmark study on the American soldier, Stouffer and colleagues during the 1940's developed a probability combination method. A more recent version of the combinatorial procedure is Winer's (1971) method of combining independent t tests. The other type of meta-analysis is the accumulation of effect sizes, such as the correlation coefficient or Cohen's d statistic. Thorndike (1933) was among the earlier researchers to accumulate results across studies using an average correlation. He also corrected the observed variance of results across studies for sampling error (unreliability). The intent of this procedure was to integrate differing research on intelligence.

While procedures for averaging correlations were available since the 1930's, as noted above, and were discussed in various behavioral statistics texts (e.g. McNemar, 1969), these procedures generally involved the use of Fisher's r to Z transformation, or were generally not used. Unfortunately no guidelines existed that allowed for a "dimensionless" statistic which could be used as a rubric or common statistic which would be independent of any specific measurement unit. Cohen (1977) developed one such statistic now in common use, the effect size statistic, or d. it was originally developed for use in statistical power analysis and to estimate the optimal sample size for a study.

In the 1970's Glass and colleagues coined the term meta-analysis, and also introduced most of the currently used procedures to psychology. Concurrently, Rosenthal was futher developing the Stouffer's combinatorial procedures. Meanwhile, Schmidt and Hunter developed what is commonly termed validity generalization procedures (Schmidt and Hunter, 1977). These involve correcting the effect sizes in the meta-analysis for sampling, and measurement error and range restriction.

Since the late 1970's the use of the quantitative review method had grown almost geometrically (Rosenthal, 1991). To give an idea of how phenomenal the growth has been, the Psych-Lit CD ROM has 909 references to this term. Before 1983, there were 51 references, while after 1982 there was 858 references.

Figure 1 shows the number of meta-analysis references published from 1975 to 1990. As shown in this figure, there is almost a geometric increase in meta-analysis related articles for the last 15 years.

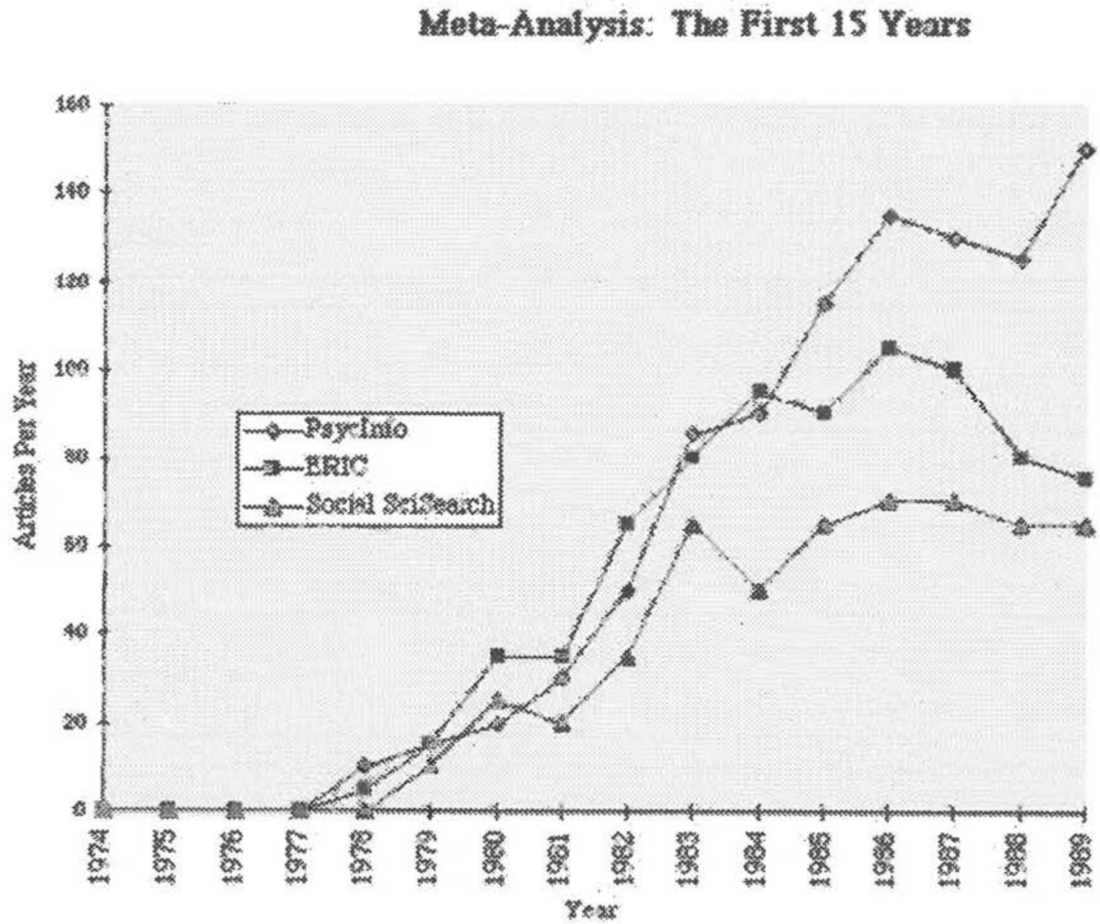

Figure 1: The First 15 years of Meta-Analysis

There are a variety of different procedures for conducting a meta-analysis involving the accumulation of correlations (r), standardized differences between mean scores (d), p values, or Z-scores (Glass, 1976, 1977; Hunter et al.,1982; Hunter and Schmidt, 1990; Rosenthal, 1991; Smith and Glass, 1977; Smith, Glass and Miller, 1980; Wolf, 1986).

Schmidt, Hunter, and their colleagues (Schmidt and Hunter, 1977; Hunter, et. al., 1982; Hunter and Schmidt, 1990) developed one method of meta-analysis that does not rely on the combination of Z-scores or probability values as the common metric. This procedure uses either r or d as the combinatorial statistic. It progressively corrects the mean r or d and their obtained variances for sampling error and then measurement error and range restriction.

In a meta-analysis the literature base is thoroughly searched for experimental and correlational studies that are relevant to the investigation. These studies become the data base for the subsequent analysis. Studies reporting on the reliability of the measures used in the various studies and their standard deviations (for range deviation adjustments) in either or both variables are included in the data base.

Converting Study Statistics to Effect Sizes

Once the data base is assembled, one converts the individual study statistic to a common metric for later accumulation (either r or d). Tables 1 and 2 show several of the more common methods of converting the individual study statistic to either r or d.

Table 1.

Formulas and Procedures for Converting Study Statistics to r.

Statistic to be Converted	Formula for Transformation to r	Notes
t	$r = \sqrt{\frac{t^2}{t^2 + df}}$	Can use with either paired or unpaired t tests
F	$r = \sqrt{\frac{F}{F + df(e)}}$	Use only with one-way ANOVAS.
Two-Way ANOVA	$r = \sqrt{\frac{(Fa * dfa)}{(Fa * dfa) + (Fb * dfb) + (Fab * dfab) + df(e)}}$	Fa = Main Effect of Interest dfa = df for A Fb = Second Main Effect dfb = df for B Fab = Interaction effect F df(ab) = Interaction df df (e) = error df
X^2	$r = \sqrt{\frac{X^2}{N}}$	n = sample size Use only when df = 1
d	$r = \frac{d}{\sqrt{d^2 + \frac{4(N-2)}{N}}}$	d = Cohen's d; N = combined sample sizes.
p	1) Convert the 2 tailed p value into a one tailed p (i.e., p/2). 2) Look up the associated Z in a normal probability table.	Can use for either exact p values or when the author reports an approximate p (e.g., p < .05).

Table 2.

Formulas and Procedures for Converting Study Statistics to d.

Statistic to be Converted	Formula for Transformation to d	Notes
Means and Standard Deviations	$d = \frac{Xe - Xc}{s_p}$	Xe Experimental Group Mean Xc Control Group Mean Sp Pooled (Within Subjects) Standard Deviation
Pooled Within Subjects Variance	$S_P^2 = \frac{(N_e - 1)\ S_e^2 + (N_c - 1)\ S_c^2}{(N_e + N_c - 2)}$	Ne Experimental Group N Nc Control Group N S2e Experimental Group Variance S2c Control Group Variance
t	$d = \frac{2t}{\sqrt{df}}$	Can use with either paired or unpaired t tests
F	$d = \frac{2\sqrt{F}}{\sqrt{df\ (error)}}$	Use only with one way ANOVAS.
r	$d = \frac{2\,r}{\sqrt{1 - r^2}}$	

Accumulating the Effect Size r and Correcting for Sample Variation

After the data are converted to a common statistic, reliabilities and range departure information are accumulated. If all the studies include reliability estimates, or range departure data, then the effect size can be corrected at the individual study level. However, most studies do not provide this information. Fortunately Hunter and Schmidt (1990) and Hunter et al (1982) provide procedures for estimating the corrections for reliability and range departures by constructing distributions for the independent and dependent variables.

When the literature base is assembled, and reliability and range information are collected, the next step is to eliminate the downwards bias caused by sampling error. This refers to the random variation due to sample size. The sample weighted mean correlation is

$$\bar{r} = \frac{\Sigma [N_i r_i]}{\Sigma N_i}$$

where $\mathbf{N_i}$ is the number of subjects in the study, and $\mathbf{r_i}$ is the effect size for the individual study. The sample weighted mean correlation is defined by the following variance formula:

$$S_r^2 = \frac{\Sigma [N_i (r_i - \bar{r})^2]}{\Sigma N_i}$$

The sample weighted mean d and its associated variance are calculated in the same manner.

Calculating and Correcting Error Variance

While the sample weighted mean correlation is not affected by sampling error, its variance is greatly increased. A two stage procedure is used to correct the variance of the sample weighted mean correlation. The first stage calculates the sampling error variance:

$$S_{e_r}^2 = \frac{K (1 - \bar{r}^2)^2}{\Sigma N_i}$$

where **K** is the number of studies in the analysis.

To estimate the biased population variance (S^2p_{xy}), uncorrected for measurement or range departure, the sampling error variance(S^2e_r) is subtracted from s^2_r .

$$S^2_{p_{xy}} = S^2_r - S^2_{e_r}$$

Correcting for Unreliability

So far this meta-analysis technique has corrected for one source of error, sampling error. There are two other forms of error, measurement error and range departure. ***Measurement error*** is assessed by measuring the impact the two reliabilities (r_{xx} and r_{yy}.) on the study results.You can do this at the study level, if reliability is reported for the X and Y variables in each and every study. Commonly in most social science research this is simply not the case. Incomplete reliability reports are far more often the norm in this area. One method around this problem is to construct a distribution of reliability coefficients, and then apply this distribution to the study results.. The reliability distribution for variable X has the mean of:

$$\bar{r}_{xx} = \frac{\Sigma\sqrt{r_{xx}}}{K}$$

where r_{xx} is the reliability for the individual study, and **K** is the number of reliability studies. The variance for this distribution is defined as:

$$S^2_{r_{xx}} = \frac{\Sigma(\sqrt{r_{xx}} - \bar{r}_{xx})^2}{K}$$

The mean reliability and the variance for the dependent variable use the same formulas.

Correcting for Range Departure

The other source of error is ***range departure***. This refers to the random deviation from rho (the estimate of the variation within the population as a whole) because of variation due to the restriction (when selecting a decreased range of scores) or inflation (when selecting extreme scores only) of the range of possible scores on any measure. Again this information is rarely reported in social science research. Generally the solution for correcting for range departure is to collect data on the standard deviations of your predictor, or X variable for as many studies in your data set as possible. Then compute the ratio of the standard deviation of the individual study to the standard deviation of some reference population (S_i/S_{ref}), "**u**". This ratio is used to construct "**c**", an estimate of the range departure for the individual study that presents the standard deviation information. The formula for calculating **c** is presented below:

$$c = \sqrt{u^2 + (1 - u^2)r^2}$$

where **u** is the ratio of the study standard deviation to the reference population standard deviation, and **r** is the effect size found in that study. Since this information is infrequently reported, a distribution of range departure elements is constructed. with a mean and variance of:

$$\overline{C} = \frac{\Sigma C_i}{\Sigma N_i} \qquad S_c^2 = \frac{\Sigma (C - \overline{C})^2}{K}$$

Estimating the Relationship within the Population

To simplify matters from here on, a notation system will be used in which the mean r_{xx} and r_{yy} will be denoted as **a** and **b** respectively and the variance of the two reliabilities will be denoted as $\underline{s}^2_a$ and $\underline{s}^2_b$.

Given these statistics, the sample weighted mean r, corrected error variance, a, b, c, and the variances for the mean reliabilities and the range departures, the relationship within the population (rho) can be estimated. **First**, correct the sample weighted mean $\underline{r}$ for measurement error and range departure using the following formula:

$$\bar{r}_{TU} = \frac{\bar{r}}{a\,b\,c}$$

Second, correct the variance of the relationship within the population for measurement error and range departure using the means and variances of the reliability and range correction factors a, b and c:

$$S^2_{r_{TU}} = \frac{S^2_{\rho_{xy}} - \bar{r}^2_{TU}(b^2\,c^2\,s^2_a + a^2\,c^2\,s^2_b + a^2\,b^2\,s^2_c)}{a^2\,b^2\,c^2}$$

This is an estimate of the variance of <u>rho</u>, or the estimate of the relationship within the population as a whole. If there is no reason to expect a serious amount of range variation across studies, as is typical with most psychological research, then the correction procedure for the range departure may be omitted (Hunter et al., 1982).

<u>**Moderator Variables**</u>

When conducting a meta-analysis, look for <u>***moderating variables***</u> (third factors that may influence the relationship of interest). Hunter et al. (1982) present a Chi-Square test for systematic variation, which is useful in determining whether there is a moderator variable present.

$$\chi^2_{K-1} = \frac{N}{(1-\bar{r}^2)^2}\,S^2_r$$

If this Chi-square value is not significant, then no moderator variable is present. Statistically this Chi Square test is very powerful, given a large enough N, it will reject the null hypothesis even if there is only trivial or meaningless variation among studies. Alternatively Hunter and Schmidt (1990) give a rule of thumb, in which the variance for the mean sample weighted r and the associated error variance (S^2_r and S^2_{er}) are compared. If the error variance accounts for less than 75% of the uncorrected variance, then a moderator variable may be present, otherwise there is no systematic variation among the studies within your data set. .

Conclusions

The overall goal of this paper was to acquaint the reader with the procedures and assumptions involved with a Hunter and Schmidt meta-analysis. Meta-Analysis provides a strong alternative to the more traditional review methods. Over the last 15 to 20 years there has been an increased criticism of the social sciences because of the increasingly confused and at times contradictory state of the research literature. While one reviewer could find a set of studies which supported his viewpoint, a second reviewer commonly found several which did not. A common conclusion in reviews was "Conflicting Results In The Literature, More Research Is Needed To Resolve This Issue." Which typically resulted in more studies which did nothing to clarify the issue. Meta-analysis offers a way out of this quagmire. By using carefully constructed and comprehensive coding and accumulation procedures, questions which cannot be easily answered with a single study can be resolved using meta-analysis.

References

Note there are several references that I need to add these are noted with an asterisk (*). There are also several references listed that while not mentioned in the text, are useful in conducting a metaanalysis.

Abramowitz, M. and Stegun, I. (1964). Handbook of mathematical functions with formulas, graphs, and mathematical tables. Washington, DC: United States Government Printing Office.

*** Cochran (1937)

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. (1st Ed). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (2nd Ed). Hillsdale, NJ: Lawrence Erlbaum Associates.

***Fisher (1932)

Cooper, H. and Hedges, L. V. (1994) The handbook of research synthesis. New York, NY: The Russel Sage Foundation.

Glass, G. (1976). Primary, secondary and meta-analysis of research. Educational Researcher, 5, 3-8.

Glass, G. (1977). Integrating findings: The meta-analysis of research. Review of Research in Education, 5, 351-379.

Haase, R. (1983). Classical and partial eta square in multifactor ANOVA designs. Educational and Psychological Measurement, 43, 35-39.

Hedges L.V. (1983) A Random Effects Model for Effect Size. Psychological Bulletin 93,388-395.

Hedges L.V. and Olkin I. (1983) Regression Models in Research Synthesis. The American Statistician, 37,137-140.

Hedges L.V. and Olkin I. (1985) Statistical Methods for Meta-Analysis. Academic Press.

Huffcutt A., Arthur, W., and Bennett W. (1993). Conducting meta-analysis using the PROC MEANS procedure in SAS. Educational and Psychological Measurement, 53, 119-131.

Hunter, J. and Schmidt, F. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Beverly Hills CA: Sage.

Hunter, J., Schmidt, F., and Jackson, G. (1982). Meta-Analysis: Cumulating research findings across studies. Beverly Hills CA: Sage.

McDaniel, M. (1986). Computer programs for calculating meta-analysis statistics. Educational and Psychological Measurement, 46, 175-177.

***McNemar, 1969

Mullen, B. and Rosenthal R. (1986). BASIC meta-analysis: Procedures and programs. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mullen, B. (1990). Advanced BASIC meta-analysis. Hillsdale, NJ: Lawrence Erlbaum Associates.

***Pearson (1933)

Rosenthal, R. (1991). Meta-analytic procedures for social research (rev ed). Beverly Hills CA: Sage.

Rosenthal R. and Rubin D. (1986) Meta-Analytic Procedures for Combining Studies with Multiple Effect Sizes. Psychological Bulletin 99,400--406.

Sacks H.S., Berrier J., Reitman D., Ancona-Berk V.A. & Chalmers T.C. (1987) Meta-analysis of randomized controlled trials. NEJM 316,450-455.

Schmidt, F. L. and Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Smith, M., and Glass, G. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32,752-760.

Smith, M. Glass, G. and Miller, T. (1980). The benefits of psychotherapy. Baltimore: The Johns Hopkins University Press.

****Stoufferet al 1940's

****Thorndike (1933)

****Tippett (1931)

Winer (1971)

Wolf, F. (1986). Meta-Analysis: Quantitative methods for research synthesis. Beverly Hills, CA: Sage.

Appendix

Conducting a Meta-Analysis, a Step by Step Guide.

1. Define the domain of research
- By independent variable
- By commonly researched variables.
- By causes and consequences of important variables.

2. Establish criteria for including studies in the review
- Published vs. unpublished study.
- The time period covered in the review.
- Operational definitions of the variables.
- The quality of a study.
- etc.

3. Determine type of effect size to use.
- Cohen's d
- Pearson's Product Moment or Point Biserial Correlation.
- Fisher's _r_ to _z_ transform

4. Search for relevant studies.
- Computer search.
- Manual search.
- Conference and Technical Symposium Presentations
- Letters to researchers in the area to be studied.

5. Select the final set of studies.
- Do individually.
- Do by more than one individual.

6. Extract data on variables of interest, sample sizes, effect sizes, reliability of measurement and other noteworthy characteristics of each study.
- Use all the data when multiple measures are reported.
- Use a subset of the data.
- Average multiple study measures to one outcome measure.

7. Code each study for characteristics that might be related to the effect size reported in the study.
- Research design factors.
- Sample Characteristics.
- Type of dependent variable.
- etc.

8. Conduct Reliability checks on the coding procedures.

- With a subset of the data, using 1 to 4 other coders.
- With all the data, using 1 to 4 other coders.

9. When there are multiple measures of independent &/or;dependent variables, decide whether to group them a priori or not.

- Theoretical diversity among variables.
- Operational measurement diversity among variables.

10. Determine the mean and variance of effect sizes across studies.

- Mean effect size weighted by sample size.
- Calculate Chi Square test for homogeneity.
- Calculate Fail Safe N.
- Between-studies variance in effect size for determining moderator variables.
- Estimation of artifactual sources of between studies variance (sampling error, attenuation due to measurement error, and/or range restriction)
- Estimation of true between-studies variance.
- Estimation of true mean effect size corrected for measurement and sampling error, and range restriction.

11. Decide whether to search for moderator variables.

- Significance Test (Chi Square test)
- Amount of between-studies variation that is artifactual.

Rule of thumb: if the variance accounted for by the error variance is less than 75% of the variance of the sample weighted correlations than there may be a moderator variable otherwise the variation is mainly due to random error (e.g., range restriction, sampling error, or measurement error).

12. Select Potential Moderators (if warranted).

- Theoretical considerations.
- Operational measurement considerations.

13. Determine the mean and variance of effect sizes within moderator subgroups.

- Procedure similar to Step 10.

Meta - Analysis:
Methods of Accumulating Results Across Research Domains
Page 23

Appendix: Steps Involved in Conducting a Meta-Analysis

1. Define the domain of research

- By independent variable
- By commonly researched variables.
- By causes and consequences of important variables.

---

2. Establish criteria for including studies in the review

- Published vs. unpublished study.
- The time period covered in the review.
- Operational definitions of the variables.
- The quality of a study.
- etc.

---

3. Determine type of effect size to use.

- Cohen's d
- Pearson's Product Moment or Point Biserial Correlation.

---

4. Search for relevant studies.

- Computer search.
- Manual search.
- Conference and Technical Symposium Presentations
- Letters to researchers in the area to be studied.

---

5. Select the final set of studies.

- Do individually.
- Do by more than one individual.

---

6. Extract data on variables of interest, sample sizes, effect sizes, reliability of measurement and other noteworthy characteristics of each study.

Note when gathering reliability and range departure information, you do not need to restrict the search to the studies used in the meta-analysis.

- Use all the data when multiple measures are reported.
- Use a subset of the data.
- Average multiple study measures to one outcome measure.

---

7. Code each study for characteristics that might be related to the effect size reported in the study.

- Research design factors.
- Sample Characteristics.
- Type of dependent variable.
- etc.

---

8. Conduct Reliability checks on the coding procedures.

- With a subset of the data, using 1 to 4 other coders.
- With all the data, using 1 to 4 other coders.

---

9. When there are multiple measures of independent &/or;dependent variables, decide whether to group them a priori or not.

- Theoretical diversity among variables.
- Operational measurement diversity among variables.

---

10. Determine the mean and variance of effect sizes across studies.

- Mean effect size weighted by sample size.
- Calculate Chi Square test for homogeneity.
- Calculate Fail Safe N.
- Between-studies variance in effect size for determining moderator variables.
- Estimation of artifactual sources of between studies variance (sampling error, attenuation due to measurement error, and/or range restriction)
- Estimation of true between-studies variance.
- Estimation of true mean effect size corrected for measurement and sampling error, and range restriction.

11. Decide whether to search for moderator variables.

- Significance Test (Chi Square test)
- Amount of between-studies variation that is artifactual.
- ***Rule of thumb***: if the variance accounted for by the error variance is less than 75% of the variance of the sample weighted correlations than there may be a moderator variable otherwise the variation is mainly due to random error (e.g., range restriction, sampling error, or measurement error.

12. Select Potential Moderators (if warranted).

- Theoretical considerations.
- Operational measurement considerations.

13. Determine the mean and variance of effect sizes within moderator subgroups --Procedure similar to Step 10.

Previous	Home	Contents	Intro	Top	Next

Southern Scholars Senior Project

Name: MARY NIKITYN Date: 12/10/03 Major: Psychology

Senior Project

A significant scholarly project, involving research, writing, or special performance, appropriate to the major in question, is ordinarily completed the senior year. The project is expected to be of sufficiently high quality to warrant a grade of A and to justify public presentation.

Under the guidance of a faculty advisor, the Senior Project should be an original work, should use primary sources when applicable, should have a table of contents and works cited page, should give convincing evidence to support a strong thesis, and should use the methods and writing style appropriate to the discipline.

The completed project, to be turned in in duplicate, must be approved by the Honors Committee in consultation with the student's supervising professor three weeks prior to graduation. Please include the advisor's name on the title page. The 2-3 hours of credit for this project is done as directed study or in a research class.

Keeping in mind the above senior project description, please describe in as much detail as you can the project you will undertake. You may attach a separate sheet if you wish:

Sheet Attached.

Signature of faculty advisor [signature] Expected date of completion 4/16/2004

Approval to be signed by faculty advisor when completed:

This project has been completed as planned: ✓

This in an "A" project: ✓

This project is worth 2-3 hours of credit: 2

Advisor's Final Signature [signature]

Chair, Honors Committee ______ Date Approved: ______

Dear Advisor, please write your final evaluation on the project on the reverse side of this page. Comment on the characteristics that make this "A" quality work.

See Separate page

A Meta-Analysis of Temperament, Personality, and Academic Achievement
Senior Project Proposal
Mary Nikityn

My senior research project will take the form of an extensive literature review and evaluation of research that addresses relationships among childhood temperament, later personality type, and life success. I will be delving into the available literature on these subjects to establish a basis for future research on possible connections between personality type and academic performance (especially math scores). In recent years, much research has been done on the relationship between gender and mathematical performance. However, there is much more variability of performance within genders than between the genders. I believe that personality type is worth investigating as another factor that might affect such academic performance and will begin working on a foundation for such research.

I will be working closely with my faculty advisor for this project, Dr. Ruth WilliamsMorris and maintaining the highest standards of academic, scientific, and research professionalism in my project. As a psychology major, I will be conforming to the American Psychological Association's guidelines for research and writing.

The project will consist of huge amounts library and internet (database) research leading to a 20 to 25-page meta-analysis of quality suitable for publication in an undergraduate research journal, as well as participation in an undergraduate conference. The research will be submitted for publication and I hope to be chosen to present it at the Dean's Luncheon in April as well.

Final Evaluation

Mary Nikityn's project is one of the best research projects that I have supervised for a senior psychology student. It is the first time that I have had the privilege to advise an undergraduate attempting a meta analysis. Mary kept regular contact with me (in person and via e-mail) and dedicated countless hours making sense of an area in social research that can be some times 'sense-less.' Meta-analyses are by definition, time consuming, requiring a grasp not only of research design, but also of statistical methodology. Analytical, writing, and mathematical skills are necessary for successful meta-analyses. Mary's project that utilizes such an approach is exemplary for undergraduate research at this level. Easily worth 3 hours of credit, this high quality project should serve Mary well as a Southern Scholar and beyond.