META-ANALYSIS OF PREDICTORS OF DENTAL SCHOOL PERFORMANCE

by

JEANETTE E. DeCASTRO

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy in Education

written under the direction of

Dr. Douglas Penfield

and approved by

_____

_____

_____

New Brunswick, New Jersey

May, 2012

ABSTRACT OF THE DISSERTATION

Meta-Analysis of Predictors of Dental School Performance

By JEANETTE E. DeCASTRO

Dissertation Chair:
Professor Douglas Penfield, Ph.D.

Accurate prediction of which candidates show the most promise of success in dental school is imperative for the candidates, the profession, and the public. Several studies suggested that predental GPAs and the Dental Admissions Test (DAT) produce a range of correlations with dental school performance measures. While there have been similarities, such as the finding by Bergman et. al., 2006 and DeBall, et. al., 2002 that the DAT's Reading Comprehension (RC) section was a significant predictor for portions of the National Board Dental Examinations Part 1 (NBDEI), correlations were disparate.

A narrative review conducted by Ranney et. al., (2005) did not apply quantitative methods and changes in the DAT and NBDE over time suggest prediction has changed as well. Ranney et. al., (2005) found extensive variation in correlations. Dental school admissions officers perceive distinctions among the GPAs of their "feeder schools," and know that grade inflation is a greater issue at some schools than others. The DAT measures students on a common rubric. Yet, correlations of the DAT with dental school performance span from low to high. The literature is unclear as to how well and how consistently the DAT and grades predict future performance for various groups (Kramer, 4/23/2012xcvii1999). Improved understanding of prediction may enhance its implementation. That is what this paper attempts to do. xcvii

A literature search found nine articles with data that could be used toward this purpose. This dissertation then applied two different methods of meta-analysis, one more streamlined, espoused by Rosenthal (1991), simply combined results, calculated confidence intervals and tested for heterogeneity of results. A second analysis followed the direction of Hunter & Schmidt (2004). After combining results, through a series of corrections, it produced correlations without the effects of range restriction and unreliability of measures.

Across both meta-analyses methods, standardized tests were more closely associated with performance measured soon thereafter. Associations between grades and criterion increased over time. Extrapolation of results to other populations is not endorsed due to shortcomings associated with available data and methodology.

Acknowledgements

I've been blessed with many advantages and these have been doubly important during my schooling. The process of writing a dissertation can get dreary, and there have been many people who brightened my path along the way. It is risky to name individuals for inevitably a key influence will be missed. I apologize in advance for omissions: all help was very much appreciated.

First there is the committee, chaired by Dr. Doug Penfield, who inherited me as an advisee from Dr. Greg Camilli, who inherited me from Dr. John Young, now at ETS, who inherited me from another advisor who left Rutgers. And so Dr. Penfield came to chair my dissertation committee after the topic he probably wouldn't have chosen himself was decided, and with far from the most promising student he ever taught, but he took me and this project on just the same. Doug, I am grateful for your patience and your generously giving of your time to thoroughly review my work and for your constructive suggestions.

Dr. Camilli taught the meta-analysis class which helped me understand the method and appreciate its potential. Completing my tier two qualifying examination with him, he gave me perhaps the most needed advice of my academic career. While I listed away all the research I wanted to do, he listened and then said, "that's all well and good Jeanette, but you can do all that later. What you need to do now is follow the most direct route to a dissertation. Here's what you need to do…" Although he soon after returned to University of Colorado, Dr. Camilli committed to see me through, and so he has. Greg, I appreciate your advice and your sticking with me, despite the distance and your other responsibilities.XCVII

Dr. Chia-Yi Chiu taught a wonderful structural equation modeling course that enabled me to more deeply understand models of prediction. Her questions about methods of test standardization ultimately strengthened that section. Chia-Yi, I am in awe of your accomplishments and appreciate your help tremendously. Dr. Joseph Holtzman, for whom being on a dissertation committee was the last thing he wanted at this stage of his career: thank you, Joe for doing it anyway and bringing content experience, perspective as former chair of a dental school admissions committee, and, humor. All the committee members shaped the work to what it is today (done!).

I appreciated all the instruction received from faculty at Rutgers, and thank Dr. Sharon Ryan for her cheerful advice and encouragement. The dean of the school where I work, Dr. Cecile Feldman, wisely advised me to choose a topic related to my work that would hold my interest and ultimately be useful in my professional life. Dr. Kim Fenesy asked throughout about progress and encouraged me to take time off for schoolwork whenever possible, acting less like a boss and more like a friend.

There were dozens of other friends at Rutgers and UMDNJ who commiserated or offered encouragement by asking how things were coming, and offered inspiration and reassurance in every encounter. Whether hallway conversations, phone calls, emails, etc., they were all soft pats on the back, saying go on, you can do this, and I thank you.

My good friends who remained even when I did not act much like a friend, not calling or seeing them for months at a time, forgetting birthdays and all sorts of things. I may have started this journey with others, but at the finish line, I see who is there.

My mother was smart and selfless enough to get a job that although did not pay well, provided me with the benefit of attvxcvii ending undergraduate college tuition free.

My father has always encouraged me by saying, “you have a good head on your shoulders, Jeanette Eileen.” [Thanks, Dad, I thought I left it somewhere else]. Each in their way told me I could be anything I wanted in life and provided a solid foundation.

There were so many family events, phone calls and visits that I didn’t make; I thank my family and extended family for their understanding. My sisters, in particular, for not only asking how things were going, but for knowing as astutely as only sisters can, when not to ask, I am grateful. I hope I can be a better daughter, sister, aunt, niece and cousin now that this is behind us and hope you know how much I have thought of you, even when actions did not seem to show it. No one could ask for a better family, and I cherish each and every one of you.

Maureen put up with my moods and lack of time to do just about anything other than work and school. She provided comfort and order; making sure I had clean clothes, paid my bills, and took walks in the sunshine. I dearly appreciated your support, and now it is time to allow me to encourage you as you pursue your own dreams.

When people say it takes a village, in my case it took much more than that: two universities, a community of colleagues, friends and family. I am grateful for all your support without which this dissertation would not have been written. Thank you.

Table of Contents

List of Tables

List of Figures

Meta-Analysis of Predictors of Dental School Performance

Chapter 1: Introduction

Prediction is a human enterprise. When societies were predominantly agrarian, weather was probably the most pressing prediction. Now that diverse occupations and competition for entry to prized careers have presented themselves, predicting performance has become, if not an obsession, a major preoccupation throughout government, business, and academia. Yet, because human behavior is remarkably complex with scores of influences, interactions, indirect and multi-directional relationships, prediction of human behavior often remains to a large degree, an unsolved puzzle. For example, most studies that attempt to foretell which applicants will be most successful in college or graduate school have nearly twice as much unexplained as explained variability (Geiser & Studley, 2002, Kuncel, Crede & Thomas, 2007, Kuncel, Hezlett & Ones, 2001, Donnon, Paolucci & Violato, 2007).

Within prediction, there are two congruent but separate aims. One is personnel selection, predicting which employees will do best in which positions, largely concerned with application of knowledge. The other is academic selection, usually concerned with mastery of subject matter, or assimilation of knowledge. Dental and medical educations represent interesting hybrid models of prediction. Whereas –to oversimplify--students in these professions typically spend the first two years of school assimilating knowledge and the last two years applying knowledge, such situations may permit fascinating comparisons between the relative successes of predictors on these two aims.

The present study, for reasons outlined below, seeks to increase understanding of estimates of predictors' effect sizes, as evidenced by Pearson product moment correlations, with regard to dental school performance.

Problem Statement

The exactness of data on which important admissions decisions are made, or factors affecting this accuracy is not fully known. Within dental education, a good number of studies have been undertaken by individuals at single institutions, some over the course of several years. Although studies have found a range of correlations between Dental Admission Test (DAT) scores and undergraduate grades as predictors, and dental school grades and National Board Dental Examinations (NBDEI and NBDEII) scores as dependent variables, to date factors have not been empirically identified that may influence the strength of these relationships. Moreover, barely any studies applied statistical methods to correct for attenuation due to restriction of range, and no studies accounted for sampling error, unreliability of measures, or other artifacts of measurement. The net result was most estimated correlations of these relationships are not as precise as possible.

This study sought to apply a wider lens that combined findings across institutions and applied additional statistical tools to try to distill truths that eluded dental educators for many years. More specifically, this study addressed three research questions:

1. What are the mean corrected correlations of DAT scores and college GPAs with dental school performance?
2. Based on available data, what variables seem to explain some of the variance in observed correlations?

3. What measurement issues inhibit uniform handling of data, and what standard reporting conventions could facilitate more exact comparisons?

Due to a rising population and increased proportion of population attending college, there is sharply increased demand for a relatively fixed number of slots at the most competitive undergraduate institutions (Alon & Tienda, 2005). The same holds true for graduate and professional education. For instance, once facing a dearth of applicants, dental schools nationally are now seeing in excess of three applicants per seat (Valachovic, 2008), but competition at the institutional level is typically higher due to multiple applications filed by each applicant.

Decisions of who gets admitted to dental school carry far-reaching repercussions for applicants, but also for schools, the profession, and the public they serve. For example, if one were to roughly calculate the economic impact on applicants, one would see that the unadjusted median net income of a general dentist practicing in the United States would be nearly four million dollars more over the course of a forty-year career as compared to the unadjusted median income of a baccalaureate-prepared applicant who was rejected and did not pursue further education. The figure for a dental specialist, such as an orthodontist, would be nearly seven and a half million more than the unadjusted median salary for a full time worker with a baccalaureate degree (calculated with information from Bureau of Labor Statistics, 2009).

Moreover, when admissions committees accept a student who is not up to the task and the student later drops out or is dismissed, he or she will typically walk away with at least $43,000 in debt per year of dental school attended (Weaver, Chmar, Hayden &

Valachovic, 2005), in addition to any existing undergraduate loans. Worse still, when a student whose skills, motivations or character are not suited for professional service graduates, the public which places its trust in dental professionals is left vulnerable to inadequate or unscrupulous treatment. These stakes are enormous but incalculable. For all these reasons, admissions decisions cannot be made lightly, and the criterion on which these decisions rest must be consistently and conscientiously evaluated and refined.

With all U.S. dental schools (Joint Commission on National Board Dental Examinations, 2009) using the Dental Admissions Test (DAT) as a factor in admission decisions, it is generally presumed that its scores adequately predict academic performance in dental school. The soundness of this presumption has received substantial scholarly attention within dental education. Empirical results have varied between schools and from year to year to the point of becoming difficult to interpret: In a review of studies of the relationship between the DAT and other predictors and performance in dental school, Ranney, Wilson, & Bennett (2005) found

> Estimates of correlation between [the DAT's] academic average (AA) and academic performance in the first year or first two years are in the range of 0.19-0.55, statistically significant in a positive direction and accounting for 4 to 30 percent of the variances in students' performances (p. 1097).

It is conventionally accepted that in most circumstances correlations of 0.1 represent small effect sizes; correlations of 0.3 suggest moderate effect sizes and correlations of 0.5 upward signify large effect sizes (Cohen, 1988). Thus, results have run the full gamut from small to large effect sizes, rendering them difficult to interpret. In the end, Ranney, Wilson, & Bennett (2005) were only able to concede that the "significant variation among studies and less than complete prediction of performance that is available" provided schools using DAT, college GPA and interviews with "defensible

methods for attempting to admit those students most likely to succeed academically and for claiming objectivity in the process (p. 1102)."

Although the Ranney, Wilson & Bennett (2005) study provided a comprehensive and informative review, and is the only one of its kind in recent history, it used studies dating back to 1965 and the DAT has changed substantially since then. Furthermore, Ranney et al., 2005 did not apply quantitative methods to synthesize findings, but instead used what has come to be known as a narrative approach. Accordingly, conclusions drawn from this review may have been more limited than possible by other means. The proposed study would update and extend the work of Ranney et al., 2005.

For all of these reasons, the purpose of this undertaking was to conduct a research synthesis of studies that analyzed the relationship between predictors and measures of dental school performance and compare results across studies to search for variables associated with effect magnitude (i.e., moderator variables) (Hall, Tickle-Degnen, Rosenthal & Mosteller, R, 1994). This aim was twofold. First, it sought to determine mean corrected correlations between predental grades and the DAT as predictors and measures of academic performance in dental school (grades and scores on licensing examinations) based on a comprehensive search, development of a database of primary studies, and the application of appropriate statistical methods. The second aim was to identify potential moderator variables that could account for at least some unexplained variance. Given the wide range of correlations reported by the Joint Commission on National Board Dental Examinations (2009) between DAT and performance across dental schools, it was hoped the study would identify the presently unknown 'rhyme or reason' as to why predictors appeared to work better at some schools than others.

At the onset, it was believed that this synthesis may have produced a more accurate picture of these relationships, enabling dental educators to assess whether the predictive validity of undergraduate GPAs and the DAT is sufficient to justify their continued role in admissions' decisions and to provide guidance as to the appropriate weights each should be afforded. It was hoped that information might have been gained for dental schools to use in determining factors that may influence the utility of the DAT. In doing so, it sought to identify variables that future researchers could incorporate into their study designs. It is essential to the survival of any field to recruit and train the most qualified personnel. Without accurate information, poor decisions could be made, and the profession could weaken.

This study had the potential to contribute to the current debate regarding the use of general standardized testing in admissions, at a time when more colleges were opting out of the SAT (Lewin, 2006). A strong relationship between DAT Survey of Natural Sciences and dental school performance indicators would support the idea that tests closely matching the topic to be mastered are capable of higher correspondence with performance indicators than generic measures of cognitive ability. If the DAT is shown to be a superior predictor than overall predental GPA, a possible explanation follows from these implications. It is possible that the quantity and type of effort required by dental school is distinct from college demands as measured by overall college GPA, and hence, prediction suffers. If predental science GPA were found to be a better predictor than predental overall GPA, it would also seem to support the supposition above that a closer match between activities assessed by predictors and outcomes resulted in a stronger relationship.

Review of Comparable Meta-Analyses

Overview. Although meta-analyses do not typically include literature reviews beyond the primary studies analyzed, it was instructive to examine meta-analyses related to predictors of graduate student achievement to inform this study with regard to customary methods and to allow a comparison of general findings. The DAT and predental grades were expected to have a predictive validity approximating that of the MCAT and pre-medical grades, and results at the graduate level provided context for interpreting the study's findings. Consequently, below is a brief outline of such articles. Table 1-1 presents a summary of correlations for relationships between the MCAT and its components with various measures of medical school performance, along with other graduate level studies.

Table 1-1
Summary of Findings of Meta-Analyses of Other Graduate School Predictors

First Author (Year)	Subjects	Predictor	Outcome	Corrected *r*
Donnon (2007)	29,701	MCAT Total	USMLE I	0.66
	27,044	"	USMLE II	0.43
	25,214	"	USMLE III	0.48
	7,419	"	Preclinical	0.43
	6,215	"	Clinical	0.39
		MCAT subtests:		
	15,508	Biology	USMLE I	0.58
	3,044	"	USMLE II	0.38
	650	"	USMLE III	0.14
	990	"	Preclinical	0.40
	275	"	Clinical	0.15
	13,568	Physical science	USMLE I	0.52
	1,384	"	USMLE II	0.28
	N/A	"	USMLE III	N/A
	990	"	Preclinical	0.26
	275	"	Clinical	0.07
	15,508	Verbal reasoning	USMLE I	0.34
	3,096	"	USMLE II	0.34
	694	"	USMLE III	0.34
	990	"	Preclinical	0.24
	275	"	Clinical	0.18

Table 1-1 *(Continued)*
Summary of Findings of Meta-Analyses of Other Graduate School Predictors

First Author (Year)	Subjects	Predictor	Outcome	Corrected *r*
Donnon	13,372	Writing Sample	USMLE I	0
	2,216	"	USMLE II	0
	N/A	"	USMLE III	N/A
	126	"	Preclinical	0
	275	"	Clinical	0
Kreiter (2007)	28,900	MCAT and premedical GPA	First and Second Year Written Exam Scores Second and Third Year Written Exam Scores	0.61
	26,752	MCAT and premedical GPA		0.58
Kuncel (2001)	14,145	GRE Verbal	Graduate Cum GPA	0.34
	45,615	"	1st Year GPA	0.34
	1,198	"	Comprehensive Exam Scores	0.44
	14,425	GRE Quantitative	Graduate Cum GPA	0.32
	45,618	"	1st Year GPA	0.38
	1,194	"	Comprehensive Exam Scores	0.26
	1,928	GRE Analytical	Graduate Cum GPA	0.36
	36,325	"	1st Year GPA	0.36
	N/A	"	Comprehensive Exam Scores	N/A
	2,413	GRE Subject	Graduate Cum GPA	0.41
	10,225	"	1st Year GPA	0.45
	534	"	Comprehensive Exam Scores	0.51
	9,748	Undergraduate GPA	Graduate Cum GPA	0.30
	42,193	"	1st Year GPA	0.33
	592	"	Comprehensive Exam Scores	0.12
Kuncel (2007)	48,915	GMAT Verbal	First –Year Graduate GPA	0.34
	48,758	GMAT Quant	First –Year Graduate GPA	0.38
	28,624	GMAT Total	First –Year Graduate GPA	0.47
	50,138	UGPA	First –Year Graduate GPA	0.35
	5,466	GMAT Verbal	Graduate GPA	0.32
	5,609	GMAT Quant	Graduate GPA	0.30
	5,201	GMAT Total	Graduate GPA	0.47
	5,609	UGPA	Graduate GPA	0.35
	1,292	Junior-Senior GPA	Graduate GPA	0.31
	680	GMAT Verbal	Persistence	0.10
	680	GMAT Quantit	Persistence	0.13
	680	GMAT Total	Persistence	0.17
	637	UGPA	Persistence	0.11

Medical College Admission Test (MCAT). In a meta-analysis of the predictive validity of the MCAT for medical school performance and medical board licensing examinations, Donnon, Paolucci, & Violato (2007) integrated the findings of 23 peer-reviewed, observational studies using a random-effects model. In addition to looking at the predictive validity of the MCAT total, their study assessed the predictive validity of the MCAT subparts (biological sciences, physical sciences, verbal reasoning, and writing sample). Outcome variables were medical school grades and/or medical board licensing exams. Due to the evolution of the MCAT over time, the study was confined to studies using MCAT scores after 1991, the last date of substantial revision to the examination, which assured more uniform sampling of predictors.

The studies sampled represented an estimated 104,912 subjects from 112 medical colleges, and included 105 correlations. For 17 of the 23 studies, correlations between MCAT scores and medical school performance outcomes were provided, permitting straightforward data extraction. For the remaining six studies, *r* was calculated from other reported data (*p*-values, beta weights, bivariate r^2 values and F ratios), reportedly using standard conversions. With regard to adjustments for artifacts of measurement, corrections for restriction of range of the independent variable (MCAT scores) were made. No corrections were made for unreliability of measures, or sampling error. Major findings were corrected medium to large predictive validity coefficient effect sizes for MCAT total with grades of 0.39 and with USMLE Step I of 0.66.

One of the shortcomings of meta-analyses that exclude non-published studies is that the included studies tend to be biased in favor of statistically significant findings,

systematically removing studies that were more likely to have found no statistically significant findings. Since the Donnon et al., (2007) study was concerned exclusively with published studies, Rosenthal's 'file-drawer' method was applied, which estimates the number of unpublished, null result studies that would be required to move the probability of a Type I error to a given significance level, in this case, 0.05. These estimates ranged from one study, for the prediction of the subtests with basic science/preclinical grades to 46 studies for the prediction of the MCAT total on the same criterion.

Kreiter & Kreiter (2007) applied a "validity generalization" (VG) perspective in a meta-analysis of MCAT scores and premedical GPA predictors with an emphasis on characterizing their prediction across years of medical training. VG is described as integrating psychometric theory with meta-analysis to interpret relationships that define validity, an approach consistent with methods applying multiple corrections for artifacts of measurement recommended by Hunter & Schmidt (2004). The Kreiter's study integrated the findings of 12 studies and incorporated the findings of a previous review for a total of 29 peer-reviewed empirical studies. Outcome variables were medical school grades, medical board licensing exams, and measures of residency and physician performance. Primary studies used MCAT scores after 1991, although results included in the previous review were older. Sample sizes ranged from 44 to 25,170, representing approximately 125,000 subjects from 112 medical colleges.

Multiple correlations of performance outcomes regressed on MCAT scores and grades were provided. Four studies provided only the correlation for MCAT scores. Since other studies showed the average additional variance explained by GPA when MCAT

was in the model was 4 percent, this was added to the results. Written exam and clinical assessment results were considered separately. Weighted mean correlations were computed directly without transformation to Z-scores, in keeping with Hunter & Schmidt's (2004) recommendations. With regard to adjustments for artifacts of measurement, the researchers reported that corrections were made for unreliability of criterion measures, and sampling error, but not for restriction of range.

Major findings were corrected mean multiple correlation coefficients for MCAT total and college grades with grades for written tests of 0.61 (uncorrected mean correlations of 0.56) in first and second years of medical school, down to corrected mean multiple correlations of 0.58 (uncorrected 0.52) in third and fourth years of medical school. Since criterion variables merged performance on USMLE and written medical school examinations, and no sample sizes were provided for the merged studies, direct comparison with the Donnon et al., (2007) study is not possible.

Large-scale studies used in the Kreiter & Kreiter (2007) and Donnon et al., (2007) studies included nearly all medical students who took national licensing examinations in a given year. There was no mention of what steps, if any were taken to ensure independent samples—leaving open the possibility that the results of some subjects were taken into account more than once by appearing in other primary studies. It is unclear therefore if some of the studies presented redundant information, providing excess weight to those effect sizes.

Graduate Record Examinations (GRE). Kuncel, Hezlett & Ones' 2001 undertaking sought to assess the validity of the GRE and undergraduate grade point average (UGPA) as predictors of graduate school performance. The GRE is used in

multiple disciplines, and thus represents a departure from the MCAT and DAT discipline-specific predictors. Like MCAT it contains a test of verbal abilities, and like the DAT assesses ability in quantitative reasoning, but also includes a separate test of analytic abilities as well as subject area knowledge for a number of fields.

Kuncel et al.'s (2001) study applied several statistical tools available to correct for artifacts of measurement. Disaggregated data were obtained from Educational Testing Service (ETS) to supplement information available in technical reports and allow for more accurate estimates of variance attributable to sampling error and differences by subject area. Hunter & Schmidt's (2004) methods were employed to correct for attenuating influences of sampling error, range restriction, and unreliability of criterion variables on the observed correlations.

The final sample included 1,521 published and unpublished studies, although methods of locating unpublished studies and the number of studies falling into the unpublished category were unreported. The authors reported eliminating less than one percent of published studies that omitted results based on significance tests and presented only results that were statistically significant. This method may have slightly reduced, but not eliminated publication bias. The selected studies represented 82,659 graduate students in 1,753 independent samples producing 6,589 correlations. Corrected correlations of GRE sections were calculated with eight different criterion (graduate GPA, first year graduate GPA, comprehensive exam scores, faculty ratings, degree attainment, time to complete, research productivity and publication citation count). Some of the outcome variables have not been well-studied in the medical and dental education studies, and therefore add little comparative value to the proposed study. With regard to grades, the

GRE subject-matter examinations yielded the strongest associations of 0.45 and 0.41 with first year and cumulative GPAs, respectively. Subject exams also had the highest correlation (0.51) with comprehensive exam scores.

Graduate Management Admission Test (GMAT). Kuncel, Crede & Thomas (2007) conducted a meta-analysis of the predictive validity of the GMAT and undergraduate GPA for academic performance in graduate business schools. The timed test consists of four sections: Quantitative (consisting of Data Sufficiency and Problem Solving), Verbal (Reading Comprehension, Critical Reasoning, and Sentence Correction), and two Analytic Writing sections (Analysis of an Issue and Analysis of an Argument). Criterion variables included first year GPA, graduating GPA, and persistence. Data were obtained from 46 studies representing nearly 50,000 test-takers.

Kuncel, Crede & Thomas (2007) applied Hunter & Schmidt-type corrections (2004), and arrived at corrected correlations for Total GMAT of 0.47, 0.47, and 0.35 with first year GPA, graduating GPA and persistence, respectively. Total GMAT was a consistently better predictor than Verbal, which was consistently better than Quantitative. Undergraduate GPA achieved corrected correlations of 0.35, 0.31, and 0.11 for the same measures, about the same as GMAT Verbal.

Oh, Schmidt, Schaffer & Le (2008) re-analyzed the Kuncel, et al. (2007) study. They disagreed with the method Kuncel, et al. (2007) used to adjust for range restriction, since adjustments for direct range restriction rather than for indirect range restriction were applied. The latter are considered more accurate when one or more variables besides the predictor under study were used in the selection process. The former are preferred when cut-off scores of the independent variable understudy are used, resulting in a

truncated sample. Oh, et al., (2008) applied adjustments for indirect range restriction and found the Kuncel, et al. (2007) findings underestimated the validity of the GMAT by approximately seven percent. For example, they found correlations between GMAT and first year GPA of 0.51, as compared to 0.47 found by Kuncel et al. (2007).

In general, the findings of other meta-analyses at the graduate level have been analogous to findings at the undergraduate level: standardized tests scores tended to predict other standardized test scores better than grades; and with the exception of the GRE, composite scores were stronger predictors than scores from subtests. Corrected correlations ranged from 0.70 for the MCAT's correlation with the USMLE Step I, to 0.10 for the GMAT Verbal's correlation with Persistence. Thus, like the DAT studies summarized by Ranney et al., (2005), relationships ran the gamut from small to large.

Conceptual Framework

Prediction of performance in dental education presents an interesting case in several theoretical areas. One being that prediction of cognitive ability is often construed as the ability to assimilate knowledge, while prediction of work performance is viewed as the ability to apply knowledge (Schmidt & Hunter 1993). Because the first two years of curriculum in dental education in the main involve assimilating knowledge, and the final two years largely involve applying that knowledge in supervised settings, this educational field presents a hybrid model, and an unusual opportunity to simultaneously assess these two types of predictions.

All involved studies employed retrospective analyses of quantitative associations. Often times, studies of standardized tests' validity are criticized for overlooking the theoretical underpinnings of constructs, and failure to start at the beginning, to review

why this test should predict student performance (Kuncel, et al., 2001). Usually the rationale behind expected prediction is not explicit. Before this meta-analysis begins to address how predictive the DAT and predental grades are, it will briefly present theoretical explanations for key constructs and relations among variables along with their descriptions.

Admission to dental school in the U.S. is largely dependent upon traditional factors such as college grade point average and scores on the standardized Dental School Admission Test (DAT). Once students have completed approximately two years of dental school, they take the National Board Dental Examinations (NBDE) Part I, the first of a two-part licensing examination; and during the fourth year of dental school, the NBDE Part II is taken. Outcome measures therefore typically include dental school GPAs, as well as performance on national board examinations.

Predictors. The Dental Admissions Test (DAT), began as the Dental Aptitude Test Battery in 1945 to assist in: decreasing student attrition (estimated at 20-25 percent of the first year class); comparing educational readiness of returning veterans' aged educational records to more recent records of non-veterans; as well as to offer a common yardstick to compare students' academic achievements, thereby offsetting the problem of differences in the meaning of grades from diverse schools (ADA, 2006).

There are currently four individual sections contained in the Dental Admission Test battery. The Survey of the Natural Sciences is a 100-item multiple choice achievement test evaluating proficiency in undergraduate coursework in basic first year biology (40 items), general chemistry (30 items), and organic chemistry (30 items). Separate scores are provided for each of the three content subtests along with a score for

"Total Science (TS)" which reports a score for the Survey of Natural Sciences as a whole. According the DAT user's manual, "while emphasis has been placed on selecting items requiring comprehension and problem solving rather than simple recall, test constructors consider the recall of information in some areas to be essential (ADA, 2006, p.2-3)."

The second section of the DAT is the Quantitative Reasoning (QR) test, consisting of 30 mathematical problems and 10 applied mathematical problems. It assumes a basic preparation level equivalent to a student beginning their first year of college. The Reading Comprehension (RC) test presents three passages of approximately 1500 words each followed by 16-17 items that examine concepts and ideas developed in the passage. The Perceptual Ability Test consists of 90 items (75 scored and 15 pre-test) presenting two and three-dimensional problems in angle block counting, paper folding, form development, and object visualization. The RC, QR, Biology (BIO), General (GC) and Organic Chemistry (OC) scores are averaged to produce a composite total score, the Academic Average (AA). Raw scores, a count of the number of correct responses, are converted to standardized scores using the Rasch psychometric model (based on the underlying log ability scale, a linear metric) to facilitate comparisons across test years (ADA, 2006). Standard scores range from one to 30. Reliability of the DAT subtests is reported as being in the range of 0.79 to 0.93 using the Kuder-Richardson Formula 20 (ADA, 2011).

The AA composite had a mean standard score of 15.53 (s. 2.27) in the base year, October 1988 (Smith, Kramer, & Kubiak, 1988), and that average increased to 17.59 (s.2.35) in 2009. In terms of distribution, over ninety percent of scores were 18 or lower in 1988, as compared to 2009 when less than sixty-six percent of scores were 18 or lower

(ADA, 2011). Thus, there is a fairly clear trend that AA scores are rising. Table 1-2 presents means and standard deviations for the 13,995 examinees in 2009.

Table 1-2
*Descriptive Statistics of Standardized Scores, Dental Admission Test, 2009 Administration**

N=13,995	Number of Items	Mean	SD
Quantitative Reasoning (QR)	40	15.66	2.79
Reading Comprehension (RC)	50	19.40	3.01
Biology (Bio)	40	17.57	2.69
General Chemistry (GC)	30	17.68	3.12
Organic Chemistry (OC)	30	17.52	3.57
Survey of Natural Sciences (TS)**	100	17.56	2.67
Perceptual Ability (PA)	90	18.17	2.92
Academic Average (AA)***	190	17.59	2.35

Note. SD=Standard Deviation
*Source: ADA, 2011, Dental Admissions Testing Program User's Manual 2009
**Combines Biology, General Chemistry and Organic Chemistry Scores
***Composite of QR, RC, TS – excludes PA

Of the 13,995 applicants who took the DAT in 2009, 58.5 percent reported their ethnicity as white, scoring on average 17.67; 27.5 percent were Asian, scoring on average 18.30; 6.4 percent were black, scoring on average 15.40, 7.2 percent were Hispanic, scoring on average 16.20; 0.4 percent were American Indian, scoring on average 16.57. More details are provided in Table 1-3. While the average score for males taking the test was 17.97, the average for females was 17.21. The existence of differences among groups' average achievement on a test by itself does not indicate that the test is less valid for some groups over others (Joint Committee on Testing Practices, 2004), it only

indicates a gap in performance that may be explained by other factors, regardless of our diligence or success in identifying those factors. There is no research within the literature on dental education that has investigated factors associated with these differences.

Table 1-3
*Demographics, Mean, Median, and Standard Deviations of DAT - Academic Average of Applicants, by Group, 2009**

Group N= 13,995	Percent of Total	AA Mean Score	AA Median Score	SD
White/Caucasian	58.5	17.67	18	2.13
Asian/Pacific Islander	27.5	18.30	18	2.42
Black/African-American	6.4	15.40	15	2.01
Hispanic/Latino	7.2	16.20	16	2.34
Native American/American Indian	0.4	16.57	16	2.34
Male	50.9	17.97	18	2.34
Female	49.1	17.21	17	2.31

Note. SD=Standard Deviation
*Source: ADA, 2011, Dental Admissions Testing Program User's Manual 2009

Generally speaking, the DAT is a timed test of cognitive ability that produces standardized scores used to compare candidates' capabilities. Primarily, however, the DAT's Survey of Natural Sciences is intended to be an achievement test that takes stock of what students have learned in several science courses (biology, general chemistry, and organic chemistry). Test content for the Survey of Natural Sciences is developed by subject matter experts with the criteria that items are appropriate, relevant, and representative of what is taught in predental courses. On the other hand, items developed for the reading comprehension section present material consistent with the type of

reading material encountered in the first year of dental school (ADA, 2006). The DAT therefore has a twofold goal: to assess whether applicants have acquired material considered requisite to comprehend information presented during dental school, as well as to assess their ability to accurately and efficiently process new scientific material. Together, these two qualities (history of mastering science subject matter, and ability to absorb new material) are believed to suggest evidence of potential for successful performance in dental school.

The common belief that the "best predictor of future success is past success" is applied, with the idea that if applicants demonstrate outstanding academic achievement in their predental science courses, they will have a tendency to perform well in dental school. While standardized tests present a 'snapshot' or a momentary sample of performance, grades and their averages are thought to represent longer-term, more comprehensive predictors since they tend to be based on multiple assessments and possibly multiple forms of assessment over time. Congruent to the DAT's Survey of the Natural Sciences, science GPA is believed to indicate prior achievement in the sciences, as well as mastery of prerequisite knowledge.

Criteria. The National Board of Dental Examinations (NBDE) are multiple-choice examinations intended to assist state boards in determining qualifications of dentists who seek licensure to practice dentistry. The Examinations are said to assess understanding of basic biomedical and dental science information and the ability to apply that information in solving problems. Part 1 is taken during or after completion of the second year of dental school and consists of 400 multiple-choice items in four topical areas that are equally and randomly distributed throughout the exam: anatomic sciences,

biochemistry-physiology, microbiology-pathology, and dental anatomy and occlusion. It contains both discipline-based items (80%) and interdisciplinary, case-based items (20%), grouped with patient scenarios. Part 2 is usually taken during the final year of dental school and consists of 500 test items: a dental discipline-based component (400 items) and case-based component (100 items). Approximately 30 percent of test items require references to the basic sciences. A single standardized score is provided for each examination, and a score of 75 or better is considered passing (ADA, 2008). Reliability statistics calculated using the Kuder-Richardson Formula 20 were reported for NBDEI of 0.92-0.95 for 2007 and 2008; for NBDEII, reliability was reported as 0.90-0.91 (Joint Commission on National Board Dental Examinations, 2009).

Since NBDE are pass/fail examinations, standardization of scores includes both standard setting and scaling/equating of scores. In conjunction with judgments about candidates' abilities, Rasch calibration statistics for criterion items are employed to set the cut score and then to equate the scores to the base year (Joint Commission on National Board Dental Examinations, 2009).

Because dentistry requires development of fine psychomotor skills, grades related to development of psychomotor skills are included in dental school GPAs, and usually have no analogue in predental GPA. The DAT Perceptual Ability (PA) is believed to predict psychomotor skills. Such skills are largely developed during the first two years of dental school, which is also when basic sciences are taught. The third and fourth year of dental school typically emphasize dental sciences and development of clinical skills through direct patient care wherein basic science and technical skills are applied.

Theoretical relationship among constructs, predictor and criterion variables. Each measure can be seen as standing in for incalculable constructs; the DAT for general cognitive abilities, perceptual ability, and prior achievement in the sciences; and grades for prior academic achievement. Board scores and dental school grades are viewed as indicators of dental school performance. With its tests of reading comprehension and quantitative abilities, the DAT bears resemblance to other instruments measuring general cognitive ability. To the extent that criterion variables may require general cognitive ability, there should be some association between the two. Namely, higher scores on measures of ability and prior achievement should be associated with higher scores on measures of performance. The DAT also assesses knowledge/achievement in the basic sciences, analogous to "job knowledge," which is believed to require general cognitive ability for its acquisition. In turn, job knowledge, is strongly related to job performance (Schmidt & Hunter, 1993), which can be viewed as analogous to dental school performance.

Measurement Issues

DAT. There are a variety of group differences reported in the literature on DAT performance and correlations involving DAT and performance indicators (e.g., ADA, 2011; Fields, Fields & Beck, 2003; Hermesch, McIntyre, Thomas & Berrong, 2005; Kingsley, Sewell, Ditmyer, O'Malley, & Galbraith, 2007). For instance, subjects' DAT scores could represent the first, second, third, or even later attempt at the examination. In the recent past, applicants could take the DAT an unlimited number of times. However, in 2007, the ADA added a provision that in order to challenge the DAT for a fourth time, students must have a dental school certify that they are applicants (ADA, 2008). Of the

nearly 14,000 applicants who took the DAT in 2009, 37.58 percent were taking it for the second or more time. Primary studies do not identify the proportion of scores as belonging to first-time test takers or repeaters. Scores for repeat testers differ significantly but marginally (from a practical standpoint) from first-time test takers (AA of 17.40, s. 2.08 vs. 17.70, s. 2.49, respectively), and this difference was slightly larger for applicants who took exam preparation courses before retaking the exam (1.1 points vs. 0.81 points for those who did not take a preparation course) (ADA, 2011). However, how well repeat scores correlate with future performance could be expected to differ from first-time testers in systematic ways. The application provided to dental schools reports the most recent exam scores above previous scores when they are available; however it is fairly commonplace for students to re-take the DAT after applications are submitted.

When a primary researcher records DAT scores for subjects, they may derive the DAT score directly from the application, from an updated report sent after the application, or may use the first, second, third or average of the DAT scores. Therefore, not knowing exactly which DAT scores were used by a researcher, or what proportion of DAT scores presented in a given study refer to scores for either first-time test takers or repeaters could mask some (minor) systematic differences in how correlations are affected.

Group differences also have been found in regard to gender (Fields, Fields & Beck, 2003; and Kingsley, Sewell, Ditmyer, O'Malley, & Galbraith, 2007). Generally speaking, scores on both predictors and outcomes tend to be slightly but significantly lower for female students. Correlations are also significantly different. Hermesch,

McIntyre, Thomas & Berrong (2005) found significant differences in predictors, performance, and associated correlations involving early acceptance and standard admissions students. Since there are great differences across schools in the proportion of students accepted via these different admissions vehicles and in the gender make up of their student populations, these differences could influence correlations, as well as homogeneity of results.

In a similar vein, according to Table 1-3, there seem to be dissimilarities among various ethnic groups' average scores on the DAT. Kramer (1999) calculated correlations between DAT scores, grades, and GPAs for a national sample of (N=8,301) students entering dental school in 1994-95. The correlation of the DAT Academic Average (AA) with overall first year dental school GPA for white/Caucasian students (N=5,086) was 0.42; for Hispanic/Latino students (N=481) 0.17; and for Black/African-American students (N=450) 0.57. This meant that only approximately three percent of the variation in overall first year dental school grades was explained by the DAT AA for Hispanic/Latino students as compared to roughly thirty three percent for Black/African-American students in that year. However, for students enrolled in 1996-97 (N=5,622), the relationships reversed, and the DAT AA correlation with overall first year grades was 0.35 for Black/African-American students (N=401), 0.41 for White/Caucasian students (N=4,734) and 0.57 for Hispanic/Latino (N=487) students. Thus, two years later, the DAT AA accounted for about twelve percent of the variation in grades for Black/African-American students and thirty three percent for Hispanic/Latino students. Correlations with some specific courses showed even more pronounced gaps and reversals over time. Because these fluctuations were "similar in magnitude," Kramer concluded "No single

set of relationships consistently favors any one group of applicants…the DAT is an unbiased predictor of dental school performance for all applicants (1999, p. 763)." However, one could easily interpret the scenario much differently, find this pattern troubling and question the validity of DAT across groups and its practical usefulness over time.

Variations in correlations associated with subject characteristics reflect only one level of variability in results, and knowing the makeup of a study population may allow researchers to more fully understand results. School-level variables, such as teaching effectiveness, faculty-to-student ratios, curricula or how closely school assessments mirror what is tested on DAT (or NBDE) can further influence correlations.

Predental GPA. It is necessary to clarify terms. While many studies refer to "undergraduate" GPA, an unknown proportion of students complete post-baccalaureate certificate programs or master's degrees before attending dental school. Grades earned while pursuing graduate work are reported on the dental application by the centralized application service, Associated American Dental Schools Application Service (AADSAS) as separate and combined values with undergraduate GPA. This could be important information since it appears that many dental school applicants who complete graduate programs do so after unsuccessfully applying for dental school. In addition, AADSAS provides GPA calculated both with and without the plusses and minuses of grades figured into the calculation, presumably because some dental schools prefer one or the other. Most authors have not defined whether the GPA used included graduate work or plusses and minuses. Throughout this proposal, candidates' grades prior to dental

school will be called “predental” so as to encompass both undergraduate and graduate records.

Students are permitted to send updated transcripts to AADSAS after the initial application, which are transmitted to the dental schools, but may or may not be used when school-based researchers collect data.

Inconsistency in how grade point averages are calculated exacerbates the problem, but has been partially alleviated by AADSAS. For instance, if a student fails a course, retakes it and earns a better grade, the failing grade might not be included in calculation of cumulative GPA (or even reported on the transcript), depending upon the practice of his/her particular institution. At some institutions both grades might figure into the GPA calculation. While averaging failing grades into the final GPA calculation would seem appropriate, it does not seem to be a uniform practice. AADSAS-- to the extent that failures appear on the transcript-- applies a uniform practice in calculating grade point averages for all applications. Therefore there is often a difference in the GPA that appears on student transcripts and the GPA reported on AADSAS application. Thus, researchers using one or the other data source will record slightly different GPAs. Until universities become more uniform in grading and reporting practices, use of predental GPAs will remain somewhat problematic, making it all the more critical for researchers to be painstaking in their described methods.

Dental School Grades. Like other institutions of higher education, practices differ among dental schools with regard to how courses for which students receive poor or failing grades are reported on the transcript, and if the GPA reported by the school

takes into account the original grade. There is no information in the literature specifically concerning reliability of dental school grades.

NBDE. Some schools engage in the practice of requiring students to pass a qualifying examination in order to be certified to take the NBDE. This would essentially remove the NBDE scores for individuals who would be expected to score the lowest. If and when previously uncertified students are later certified to take NBDE, the student may have undergone a review process made more intense by the existence of a qualifying examination. It could be expected that the correlations from schools that adhere to this practice would be systematically different from those that do not.

Application of the Rasch model (which is used to standardize both the DAT and NBDEs) to multiple-choice tests has been shown to be inaccurate (Divgi, 2005) and is considered controversial because the model assumes that items cannot be answered correctly through guessing (Zwick, Thayer & Wingersky, 1994). How these potential inaccuracies affect consistency of scores across administrations and relationships of DAT and NBDE scores with each other or other variables under study is unknown. However, Kramer & DeMarais (1992) found very little difference in standards and failure rates when the Rasch method was compared to a norm-referenced approach on NBDE II. Moreover, because KR-20 estimates do not address transient error, it is likely that reliability statistics for DAT and NBDE were overstated (Hunter & Schmidt, 2004).

Theoretically, due to the many attributes believed to be related to dental school performance, including cognitive abilities, psychomotor, time management, and study skills as well as motivation, work-ethic and professionalism, any single test is unable to measure all relevant characteristics. Others have held that quality of dental school

instruction and students' backgrounds may influence dental school performance (Potter & McDonald, 1985). It has already been noted that predental as well as dental school GPAs have their shortcomings. Statistically, there is not a wide range of variability among applicants selected to attend dental school, and once there, their array of performance is relatively narrow. This situation creates a restricted range for analysis and in turn, attenuated correlations. Methodologically, combining the effect size of studies conducted with diverse methods and variables presents its own challenges. In summary, theoretical, statistical and methodological issues will be taken into consideration throughout the proposed study.

Chapter 2: Methodology

Study Retrieval

Search mechanisms. Electronic databases, including Psychological Abstracts, Medline, PubMed, Academic Search Premier, Searchlight, Dissertation Abstracts, and PsycLIT, PsychINFO were systematically searched to find studies. In addition, the Journal of Dental Education, the most common publication in which studies of this topic are found, was searched by hand. Technical reports from the American Dental Association (ADA), publisher of the DAT, NBDE Part 1, and Part 2, were reviewed to find related studies not presented via the electronic search. Citations listed in all retrieved works were examined to identify additional studies. Conversely, studies that reference retrieved works were reviewed.

To reduce publication bias, efforts were made to locate relevant unpublished research. Emails seeking information concerning ongoing or unpublished projects were sent to individuals in organizations that supported related research, such as the National Institute of Dental and Craniofacial Research's Division of Extramural Research -Social and Behavioral Health Branch, the American Dental Association (ADA), and the American Dental Education Association. Private foundations, such as the Robert Wood Johnson foundation, that have funded research in oral health education and their list of funded projects were searched. In addition, contacts were made with key individuals within dental education, including those who published on this subject, with the goal of unearthing unpublished studies.

Selection criteria. Only studies that met the following criteria were included in the final report:

1. Based on class years included in the study, it can be assumed that the majority of subjects were tested using the current version of the DAT (administered Spring, 1990 or later) or its subtests
2. Presented observed findings of DAT scores or predental grades related to dental school grades and/or national licensing examinations
3. Were published (or unpublished) in English
4. Relied on data collected at American dental schools.

Inasmuch as the DAT underwent major revisions to test content, number of items in certain sections, format, and the method by which standardized scores were calculated over the years (ADA, 2006), this study excluded studies that relied predominately on DAT data collected prior to Spring, 1990, the last date for which noteworthy changes were reported in the *ADA User's Manual* (ADA, 2011). Due to differences in DAT test components in U.S. and Canada, only studies from American dental schools were used. By restricting the sample to works reported in English involving students enrolled at U.S. dental schools, more consistent sampling was possible.

Care was taken not to include duplicate samples by including school and class years in the coding protocol. Publications that would have duplicated samples already reported were excluded. In the case of Sandow et. al, (2002) and Behar-Hornstein et. al, (2011), both studies were conducted with samples drawn from the University of Florida, but for non-overlapping class years, so neither was excluded.

Rosenthal (1995) advises that a meta-analysis can be conducted with as few as two studies, but too few would produce relatively unstable results. He recommends when there is an insufficient quantity of studies available that the meta-analysis be incorporated

as an extension of the results section in an additional study. Given the relatively narrow inclusion criteria, it was decided that if fewer than five studies were identified as meeting the criteria, the meta-analytic results would be incorporated into the results section of a recently conducted study (DeCastro, 2010) that has yet to be submitted for publication.

Coding. A coding protocol was developed *a priori* to assist in recording essential information from each study, including title, authors, author's primary affiliation, year of publication, publication, study design, measures of predictors, measures of performance, obtained estimates (means, standard deviations, correlations, sample size, *p* values, sample description if any, school, class years involved in study, and number of class years included in analysis. Institutional-level variables, publicly available in *2010 ADEA Official Guide to Dental School Applicants* (American Dental Education Association, 2010) and school websites, were recorded which included course-work required for admissions, average and range of incoming classes' DAT scores and GPAs, entering class size, school curriculum type, prerequisites, faculty: student ratio, applicant: enrolled ratio, type of funding (state, private, quasi-private); demographic enrollment statistics; and geographic description (rural, urban, suburban). Characteristics were added when literature, patterns, findings or hunches seemed to warrant their inclusion. Clarification of some descriptors and additional information on school-level variables at the time of the primary study or more precise descriptions of curriculum was sought through direct contact with individual schools/authors, ADA and ADEA. All coding was performed by the author, which controlled inter-rater reliability.

Statistical Methods

General Framework. Meta-analyses vary in their intricacy not only due to the studies involved but also according to the general framework used. Rosenthal (1995) suggests that these can be divided into the more detailed and quantitatively demanding, such as Glass (1980), Hedges & Olkin (1985) and Hunter & Schmidt (2004); or, more basic systems described by Rosenthal (1991), and Cooper (2010). To begin, the study closely followed suggested guidelines in method and format outlined by Rosenthal (1991 and 1995), supplemented by Cooper (2010) and others as necessary. In parallel, however, procedures advocated by Hunter & Schmidt (2004) were conducted and results were compared. For example, whereas Rosenthal advocated the use of *r* transformed to *Z*, and Hunter & Schmidt do not, comparison of uncorrected results may be of interest. Further, as Rosenthal (1991) aptly points out the types of corrections recommended by Hunter & Schmidt are intended to produce correlations that would exist in a perfect world, and as such, these indices would hold little practical value; there will always be imperfections in measures, sampling and other methods.

Although the Rosenthal (1991) method presents one of the most straightforward methods for combining results, it does not correct for error and bias in research findings to the degree of other methods such as Hunter & Schmidt (2004). Corrections for sampling error, restriction of range, and measurement error produce a correlation with credibility intervals suggested to be more representative of a relationship at the construct level; this provides information concerning hypothetical, rather than operational relationships.

At least some variation between studies is likely due to artifacts of measurement. Hunter and Schmidt (2004) suggest a multitude of corrections for artifacts of

measurement. Systematic artifacts of measurement include differences in the reliability of dependent variables, such as NBDE Part I scores, or first year grades. Since the dependent variables were not combined but were assessed separately, there were no obvious inter-study differences in reliability. However, adjustments were made to correct for unreliability of individual measures.

Model. Conceptually, the fixed effects model suggests that there is a single population effect size, theta, and all differences between studies are due to sampling error. That is, each study measured a different subset of the population and therefore had different results. The random effects model, on the other hand, suggests that there is a distribution of population effect sizes, rather than a single population effect size. Differences between studies are therefore believed to be due, at least in part, to real differences in the underlying population (Shadish & Haddok, 1994). In addition, studies, not individuals within studies serve as the sampling unit, a more conservative approach albeit with lower power. Rosenthal (1995) reminds researchers not to get too 'hung up' on the random versus fixed effect issue, because

> …there is precious little random sampling of studies in meta-analytic work. Indeed, even in the fixed effects model, when one generalizes to other sampling units within the studies one assumes that the new sampling units will be randomly sampled within the study from the same population from which one sampled the original sampling units. However, it is very seldom that in behavioral or biomedical research that one samples participants or patients randomly. Hence 'random' should be thought of as quasi-random at best (p. 187).

Therefore, while fixed effects methods were applied for Rosenthal-type analyses, random effects tests of significance were used in order to generalize results to other studies from the same population of studies from which the retrieved studies were sampled (Rosenthal, 1995). Nevertheless, the emphasis of this study has been on

ascertaining the magnitude of the mean effect size rather than its probability/significance. Power becomes great enough to detect even meaningless effect sizes in meta-analyses that combine significance levels (Cooper, 2010). On the same basis, vote counting was not employed.

Wherever viable, when the analytical process approached a decision point for which there was not a reasonable amount of consensus, decisions were made on conceptual, statistical, and practical (i.e., the availability of data) bases, in that priority order. In the event future researchers wish to compare results to those presently under discussion but have concluded (based on currently unavailable information) that another approach would have produced more accurate results, alternative methods and results are presented to the extent feasible.

Overview: Sequence of steps. The Rosenthal-style meta-analysis was implemented in six main stages. First, the observed means, sample size, variances and correlations were recorded for each study. Works reporting relationships other than correlations, such as t, F, or *p* recorded that data. Second, the correlations (or other data) were transformed to Z-scores. Third the Z-scores were weighted by their sample sizes. Fourth, the Z-scores were averaged across studies and confidence intervals were calculated. Fifth, the cumulative Z-score and confidence intervals were transformed back to correlations. Finally, homogeneity and significance of results were tested.

The Hunter-Schmidt style meta-analysis was accomplished in four stages. First, means and variances of observed correlations were collected from each study. Second, variances of correlations were corrected for sampling error. Third, the correlations themselves underwent corrections for measurement error and range variation. Corrections

were applied across studies based on estimates and formulae. The decision to apply corrections to studies as a group was based on availability of data. For example, sample size was available for all studies, and as this is the primary data needed to estimate sampling error, it was calculated at the individual level. However, artifact distribution methods were used to correct for reliability and range restriction. Finally, as indicated, data were analyzed by subsets to test for moderator variables.

Methods to combine results. DAT and college grades were the independent variables; dependent variables were dental school performance indicated by dental school grades and performance on licensing examinations. Pearson's product-moment correlations (*r*) were used as the measure of effect size. This choice was appropriate because the both independent and dependent variables were continuous (Cooper, 2010). For the eight (88 percent of) studies reporting bivariate correlations between DAT scores and college grades with dental school performance outcomes, correlations were directly entered into the database for later transformation. In the Rosenthal-method, transformation of *r* to Z corrected for bias in the *r*-distribution. This transformation was performed by automated calculations available in statistical software.

Authors of primary studies reporting only results from multiple regressions were contacted to determine if bivariate correlations were available. Data was available for Bergman et al., 2006.

After each *r* index was transformed to its corresponding *Z* score, the average value of *Z* was calculated by weighting the obtained Z_s by their corresponding sample size, or more precisely by N-3, as recommended by Rosenthal (1991). Because larger samples yield more precise estimates, this is a generally accepted practice (Cooper,

2010). Ninety-five percent confidence intervals were computed around the mean effect size using a random effects model. This entailed computing the standard error of the mean *Z* and using that in turn to compute confidence intervals, which were applied to the *Z* prior to transformation back to *r*.

Some meta-analysts argue against using more than one effect size from a study. Using more than one effect size from each study does not violate independence as others have suggested, because the samples are not added into analysis of the same effect size twice (Rosenthal, 1991).

Hunter & Schmidt (2004) make a case that because Fisher's *r* to Z transformation applies more weight to large correlations than to small ones, it causes an upward bias, resulting in correlations larger by about 0.03. Consequently, in the parallel analysis, sample-size weighted-average calculations were computed directly using individual Pearson's product-moment correlations (*r*). The corresponding variance across studies was calculated as a frequency-weighted average squared error, which gave greater weight to large studies than to small studies. All data was collected and analyzed using SPSS© (version 19, 2010, Biostat Inc., Englewood, NJ). Some graphs were produced using Microsoft Excel and Microsoft Powerpoint.

Significance testing. The Stouffer method as described by Rosenthal (1991) was used to test the significance of the resulting estimates of mean effect sizes. The standard normal deviate (*Z*) associated with each *p* value was computed and averaged to test the overall result. This is a fixed effect method, which would permit generalizations only to studies included within the review. To expand generalizability to other studies from the same population from which the retrieved studies were drawn, and as recommended by

Rosenthal (1991), a random effects approach was used by means of a simple t-test applied to the mean derived *Z*.

Measuring sampling error. Estimation of sampling error was accomplished by application of a statistical formula using the mean correlation to derive the average of the sampling error variances within studies (Hunter & Schmidt, 2004). These results permitted calculation of estimated percent of variation in observed correlations due to sampling error. Sampling error was removed from error variance estimates.

Corrections for measurement error. According to Hunter & Schmidt (2004), when range restriction is indirect, reliability corrections must be made prior to the range restriction correction using reliability values for the restricted group (Hunter & Schmidt, 2004). Therefore, corrections for measurement error were introduced before corrections for restriction of range.

The same standardized tests, the DAT and NBDE were used across studies, making varying reliability of measures among studies not as much of a concern as imperfect measurement from standard instruments affecting all studies. Since primary studies in this meta-analysis did not report information on the reliability of examined variables, artifact distribution methods described by Hunter & Schmidt (2004) were used. Various data sources and formulae were used to construct separate reliability distributions for each predictor and criterion variable. Variance in DAT test scores and predental GPAs in the applicant and incumbent populations were obtained from the American Dental Education Association (ADEA)'s Center for Educational Policy Research via a custom data request. Reliability of incumbent licensing exam scores was available in ADA-produced technical reports, but reliability for applicant scores on

NBDEs (dependent variables) were estimated by formula, since applicant performance data was not available. Reliability of the DAT AA score was estimated since ADA did not compute reliability statistics for the composite score (Tsai, 2011).

Reliability information was less available with regard to applicant and dental school grades. Reliability estimates for college GPAs were developed by reviewing published estimates of the reliability of GPA (Barritt, 1966, Reilly & Warech, 1993, Stricker, Rock, Burton, Muraki & Jierele, 1994, and Young, 1988). Kreiter & Kreiter (2007) estimated reliability of first-year medical school grades (which was the same as the average estimated reliability of college grades) and this served as a proxy to an estimate of the reliability of dental school grades. Applicant and incumbent reliability was computed by formulae provided in Hunter & Schmidt, 2004. Consequently, using available and computed estimates of reliability, standard procedures to correct for measurement error in independent and dependent variables were applied to mean correlations. Although as a practical matter measurement error in the predictor is usually not corrected, Hunter & Schmidt (2004) advise that both variables must be corrected before applying indirect range restriction formulae.

Adjustments for range restriction and attrition artifacts. Correlations corrected for unreliability were then corrected for indirect range restriction, using formulae provided by Hunter & Schmidt (2004). Variances, on the other hand, were corrected first for restriction of range, and then for reliability. Corrected credibility intervals were computed using Taylor's series (Hunter & Schmidt, 2004).

Range restriction in the dependent variable produced modest attrition artifacts. Specifically, poor performing students left school prematurely; removing the lowest

performing students from the data pool. The resulting population differed from the pool of all enrolled students and applicants due to attrition effects. Within dental education, this effect is believed to be relatively slight, since retention in dental schools appears to average at least ninety percent.

On the other hand, the selection effect of the DAT and predental GPAs greatly reduced the range of the independent variable; namely, applicants with extremely low DAT scores and or predental grades typically would not be accepted (and may have been discouraged from applying at all) and this pool differs from the pool of all applicants. In terms of size, the pool is reduced by approximately two-thirds. "Indirect" range restriction is said to be operational when other factors besides the independent variable are considered in the selection process. Hunter & Schmidt (2004) advised that if the standard deviation of the independent variable was stable across studies, correction for range restriction is unnecessary. However, since differences were evident corrections were applied.

Identifying study characteristics and potential moderators. It was decided *a priori* that if unaccounted variance was found to be greater than 20 percent that study characteristics would be reviewed to determine patterns among study or school features. Once effect sizes were integrated into a mean effect size, the variance in effect sizes across findings was analyzed. Heterogeneity of effect size estimates were assessed to assist in identifying significant differences that may suggest all the effect sizes were not drawn from the same population. Namely, it determined how likely it is that the displayed variance of effect sizes is due to sampling error. Calculation of a Q_t statistic is a common practice and recommended by Cooper (2010).

As recommended by Rosenthal & DiMatteo (2001), examination of variability in effect sizes and their standard deviations were conducted informally with graphs and charts since significance tests are highly influenced by sample size. Degrees of variability were assessed by comparing the ratio of standard deviations to scores, and plotting them to identify similar groupings. Information from the database of school characteristics were graphed and assessed.

Rather than conducting a formal test of heterogeneity of variance, Hunter & Schmidt (2004) recommend determining if the observed variances in effect sizes were twice as large as expected, and if so, that it be considered reliably different.

In addition, the ADA issues periodic validity reports about the DAT's prediction of first and second year grades for all dental schools, based on information provided by the schools' registrars. The grades are broken down into overall, biomedical sciences and technique courses. While this information would have been valuable, the ADA does not identify individual schools in the report, single school codes are given only to the school's administration. Although the ADA was contacted to request release of information that would be helpful in identifying patterns, it did not feel it was at liberty to release this information (Tsai, 2011).

Publication and related bias. Publication bias refers to the higher probability of publication of studies finding statistically significant treatment effects. "Small-study effects" occur when smaller studies in a meta-analysis show larger treatment effects because larger treatment effects are required to achieve significant results when the sample size is smaller (Sterne, Gavaghan & Egger, 2000). As recommended, tests for small-study effect size and publication bias were performed. Plots of effect size estimates

against sample size were used to detect such biases as described by Egger, Smith, Schneider & Minder (1997).

Interpretation of results. As noted earlier, Rosenthal (1995) suggests the computation of confidence intervals around the calculated estimated mean correlation, and significance testing through the use of a t-test to determine generalizability to other studies. Hunter & Schmidt (2004) emphasize the use of credibility intervals when interpreting meta-analytic results, holding they are particularly appropriate in the context of random-effects models, which allow for the possibility of variation in parameters across studies. Credibility intervals estimate the range of real differences after accounting for sampling error. Whereas confidence intervals express the likely amount of error in the estimate of the mean value of *p* due to sampling error, credibility intervals refer to the distribution of parameter values, formed by the use of SD_p rather than the standard error of the mean of *p* (since it is held that variance due to sampling error has been removed from the estimate of SD_p).

Reporting results. A combination of the reporting conventions advocated by Halversen (1994), Rosenthal (1995) and Cooper (2010) was used to guide the format of the final report's major subject headings and content.

In sum, then, by following these procedures a meta-analysis was conducted that produced estimates of mean correlations, confidence intervals and tests of significance; corrected mean correlations with credibility intervals, and, to a limited extent, identification of possible moderator variables.

Hypothesis

This study began with a hunch that DAT and past college performance are better predictors than the credit they have received. It was expected that corrected correlations would be found that were a good deal higher than those reported which are markedly attenuated. I hypothesized that the meta-analysis would find average effect sizes in the high-medium to large range for the DAT and predental grades overall. Without much hope of obtaining individual student demographics with available data, I expected to see some relationship between school-level variables and the effect size of the DAT.

Chapter 3: Results

Search results

The opening search produced 15 studies; 9 were found to meet inclusion criteria, and 6 were found not to meet one or more of the four criteria. For any given pairing of independent and dependent variables, there were at most six studies reporting correlations. These journal articles are listed in Table 3-1 (Appendix A), along with means and standard deviations of reported DAT scores and dental school performance measures. Over the course of this project, the search and statistical methods employed were repeated to ensure any new studies and their results were included.

By and large, primary studies relied on retrospective cohort designs reporting bivariate correlations, although some studies reported results of multiple regression analyses. Specifically, for seven (78 percent) of the nine studies extracted, correlation(s) between the DAT scores and/or predental grades and dental school performance outcomes were provided in the results. For one study (Bergman, et al., 2006), however, the authors provided the raw data needed to convert to correlations. Fields, Fields & Beck (2003), provided data as well since the correlations of interest were not published. Several studies were eliminated because the correlations of interest were not reported and raw data was either destroyed or otherwise unavailable.

Results observed in primary studies

Retrieved studies were markedly similar in operational definitions. Independent variables uniformly included predental school GPAs and/or DAT scores. Dependent variables were consistently dental school grades and NBDE scores. A few studies tested

additional relationships and presented correlations as comparative data for subgroups and group data was therefore combined for analysis.

Descriptive Statistics. Results for nine studies conducted at eight separate schools represented 2,853 students applying to dental school between 1991and 2010. Table 3-1 (Appendix A) provides mean scores and standard deviations from the primary studies. Reported scores and grades are essentially equivalent to or higher than the national medians reported in Table 1-2, and no scores reported were more than a few tenths lower than the reported national medians for that year. There was less variability among scores and grades in most primary studies than reported nationally.

The rudiments of associations between predictors and performance indicators can be previewed here: studies from schools with higher mean DAT scores or predental GPAs tended to display higher mean performance on National Board examinations and slightly less variability relative to other schools. Conversely, it appeared that schools with lower mean incoming scores had correspondingly lower mean criterion values and slightly more variability.

Figure 3-1: Graph of NBDEI Scores as a function of AA Means

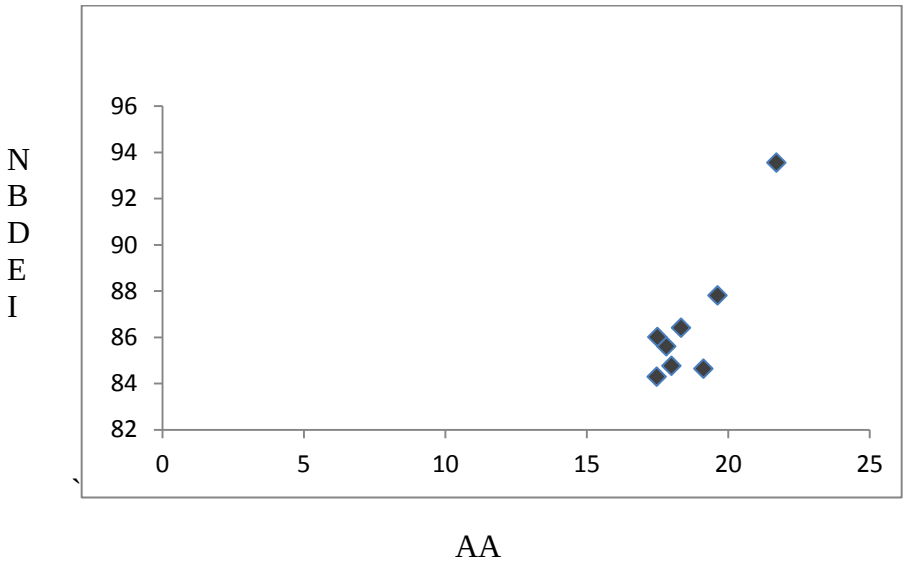


Figure 3-1 displays what appears to be a steep positive relationship between AA and NBDEI scores found in primary studies. Correlations found in the primary studies are reported below. Table 3-2 presents correlations involving overall predental GPA (OAGPA), predental science GPA (SCIGPA) and dental school performance indicators, such as first year dental school GPA (YR1GPA), and cumulative fourth year dental school GPA (YR4GPA).

Table 3-2

Correlations of Predental GPAs and Dental School Performance Indicators

First author:	Curtis	DeCastro	Fields	Hermesh	Holmes	Kingsley	Sandow	Range
Sample size:	49	345	451	361	566	210	410	49-566
Variable pair	Reported Correlations							Range
OAGPA: YR1GPA	0.210	0.189	No data	No data	No data	0.087	No data	0.087-0.210
OAGPA: YR4GPA	0.330	No data	0.451	0.429	0.529	No data	No data	0.330-0.529
OAGPA: NBDEI	No data	0.064	0.349	0.290	0.497	0.222	No data	0.064-0.497
OAGPA: NBDEII	No data	No data	0.318	0.307	0.433	No data	No data	0.307-0.433
SCIGPA: YR1GPA	0.270	0.223	No data	No data	No data	0.068	0.413	0.068-0.413
SCIGPA: YR4GPA	0.320	No data	0.455	No data	0.537	No data	0.425	0.320-0.537
SCIGPA: NBDEI	No data	0.156	0.344	No data	0.527	0.227	0.309	0.156-0.527
SCIGPA: NBDEII	No data	No data	0.340	No data	No data	0.460	0.280	0.280-0.460

Table 3-3

Correlations of DAT Scores and Dental School Performance Indicators

First author:	Behar-Hornstein	Bergman	Curtis	DeCastro	Fields	Holmes	Sandow	Range
Sample size:	209	249	49	351	451	566	410	49-566
Variable pair	Reported Correlations							Range
AA:YR1GPA	No data	No data	0.36	0.376	No data	No data	0.475	0.36-0.475
AA:YR4GPA	No data	No data	0.28	No data	0.272	0.494	0.317	0.272-0.494
AA:NBDEI	0.450	0.280	No data	0.325	0.419	0.610	0.507	0.280-0.610
AA:NBDEII	0.270	0.229	No data	0.291	0.305	0.524	0.433	0.229-0.524
TS:NBDEI	0.420	0.202	No data	0.313	0.233	0.582	No data	0.202-0.582
TS:NBDEII	0.180	0.206	No data	0.082	0.427	0.469	No data	0.082-0.469
PA: YR1GPA	No data	No data	0.050	0.230	No data	No data	0.279	0.05-0.279
PA: YR4GPA	No data	No data	0.030	No data	0.133	0.370	0.198	0.030-0.370
PA:NBDEI	0.060	0.111	No data	0.142	0.120	0.363	0.263	0.060-0.363
PA:NBDEII	0.130	0.036	No data	0.142	0.163	0.344	0.304	0.036-0.344

Table 3-3 above provides correlations among DAT scores and dental school performance indicators. DAT scores include Academic Average (AA), Survey of Natural Sciences (TS), and Perceptual Ability (PA). Among DAT scores, PA usually produced smaller correlations, and AA produced among the largest, but only slightly larger than TS. Among predental grades, the SCIGPA generally produced slightly stronger correlations than overall GPA.

Pooled results and homogeneity of results

Rosenthal-method results. Along with total sample size and number of studies, results of transforming and combining correlations to produce an average correlation using the Rosenthal-recommended methods are presented in table 3-4, as are results of tests of homogeneity.

Table 3-4 (Rosenthal-Style Results)
Combined Data by Variable Pair: Sample Sizes, Average r, Probability of Correlation, 95 % Confidence Interval and Results of Q_t Test

Variable Pair	N	# Of Studies	Avg. *r*	Probability	95 % Conf. Int.	Significantly Different Correlations? (Q_t)
AA:Yr1GPA	810	3	0.425	<.0001	0.367 - 0.480	no
AA:Yr4GPA	1519	4	0.373	<.0001	0.328 - 0.416	yes
AA:NBDEI	1763	6	0.470	<.0001	0.436 - 0.502	yes
AA:NBDEII	1409	6	0.379	<.0001	0.342 - 0.415	yes
PA:Yr1GPA	801	3	0.186	<.0001	0.118 - 0.252	No
PA:Yr4GPA	1476	4	0.195	<.0001	0.146 - 0.244	Yes
PA:NBDE1	2233	6	0.207	<.0001	0.167 - 0.246	Yes
PA:NBDEII	1544	6	0.195	<.0001	0.155 - 0.235	Yes
TS:NBDE1	1713	4	0.389	<.0001	0.348 - 0.429	Yes
TS:NBDEII	1585	3	0.352	<.0001	0.308 - 0.394	Yes
OAGPA:Yr1GPA	622	3	0.150	<.0001	0.070 - 0.228	No
OAGPA:Yr4GPA	1435	4	0.444	<.0001	0.431 - 0.511	No
OAGPA:NBDEI	1844	5	0.325	<.0001	0.283 - 0.365	Yes
OAGPA:NBDEII	1386	3	0.333	<.0001	0.271 - 0.393	Yes
SCIGPA:Yr1GPA	1032	3	0.324	<.0001	0.261 - 0.385	Yes

SCIGPA:Yr4GPA	1479	5	0.432	<.0001	0.392 - 0.470	Yes
SCIGPA:NBDEI	1438	4	0.329	<.0001	0.289 - 0.368	Yes
SCIGPA:NBDEII	1427	3	0.381	<.0001	0.336 - 0.424	Yes

Note. N=combined sample size

Without corrections, when Rosenthal-style methods were applied, based on 1,435 subjects across four studies, OAGPA had the highest correlation with YR4GPA, marginally greater than the correlation with SCIGPA. Each explained about 22 percent of variance in YR4GPA. Based on 1,519 subjects across 4 studies, the association of DAT AA and fourth year GPA explained nearly 14 percent of variance.

The correlation for PA with NBDEI represented the largest amount of subjects, 2,233, across six studies. While PA was found to have significant correlations with all performance variables, most correlations associated with PA were amongst the lowest; it explained only between four and six percent of variance of any of the dependent variables.

In summary, before corrections, data in table 3-4 indicate the mean uncorrected correlations ranged from 0.15 for OAGPA with YR1GPA to 0.47 for AA and NBDEI, as well as SCIGPA and OAGPA (individually) with YR4 GPA. Significant variability was found among most of the 18 correlations. Results of the Q-tests for homogeneity of variance are also presented in table 3-4 and show significant heterogeneity among the variances. For the smallest sample size among the correlations (622), a correlation of approximately .08 or larger was required to reach .05 significance. Accordingly, all correlations were significant at the .05 level.

Hunter & Schmidt-method results. Up to the third column in table 3-5 displays the end products of what Hunter & Schmidt (2004) call a "bare-bones" meta-analysis --one that accounts only for sampling error—that is, a mean correlation coefficient, corresponding variance and estimate of the amount of variance in correlations due to sampling error. After removing error variance from the variance of observed correlations, the residual variance (not displayed) which estimated population variance was quite small, ranging from 0.001 for AA:NBDEI to 0.018 for TS:NBDEII.

Weighted bivariate correlations produced using Hunter & Schmidt methods (2004) without the *r*-to-*z* transformation were essentially the same as those produced with the transformation, but as expected slightly smaller for larger correlations. The correlation involving PA and first year GPA had the highest amount of variation (73 percent) attributed to sampling error. Conversely, correlation of TS and NBDEI had little variance (about five percent) accounted for by sampling error. In fact, correlations with NBDE I or II as dependent variable had no more than 17 percent of variation due to sampling error. The standard errors corrected using Taylors' series (Hunter & Schmidt, 2004) reflect sampling error in proportion to the overall correction applied to the correlations. The mean weighted correlations (before corrections), corrected average standard error, sampling error and correlations corrected for first reliability in the independent and dependent variables and then both reliability and range restriction using Hunter & Schmidt methods (2004), are reported in table 3-5.

Table 3-5

Mean weighted correlations, corrected average standard error, sampling error and correlations corrected for reliability and range restriction

Variable Pair r	Mean weighted *r*	Standard Error	% sampling error	Corrected for reliability in x and y	Corrected for range restriction
AA:YR1GPA	0.427	0.036	64	0.526	0.561
AA:YR4GPA	0.371	0.028	15	0.441	0.474
AA:NBDEI	0.464	0.019	11	0.491	0.525
AA:NBDEII	0.379	0.021	9	0.411	0.443
PA:YR1GPA	0.245	0.043	73	0.308	0.322
PA:YR4GPA	0.196	0.023	53	0.238	0.249
PA:NBDEI	0.205	0.023	17	0.221	0.232
PA:NBDEII	0.222	0.022	11	0.246	0.258
TS:NBDEI	0.379	0.024	5	0.404	0.415
TS:NBDEII	0.346	0.025	10	0.379	0.389
OAGPA:YR1GPA	0.150	0.073	69	0.230	0.280
OAGPA:NBDEI	0.332	0.031	7	0.438	0.516
OAGPA:NBDEII	0.329	0.047	7	0.446	0.524
OAGPA:YR4GPA	0.473	0.028	48	0.703	0.774
SCIGPA:YR1GPA	0.268	0.064	19	0.420	0.564
SCIGPA:YR4GPA	0.475	0.026	40	0.720	0.837
SCIGPA:NBDEI	0.356	0.030	10	0.479	0.627
SCIGPA:NBDEII	0.337	0.038	6	0.466	0.613

Table 3-6 provides information concerning the percentage of variance explained by the corrected correlations and the 95 percent credibility interval. Overall GPA predicted about 8 percent of variance in first year dental school GPA, but approximately 60 percent of fourth year dental school GPA. SCIGPA explained 38 and 78 percent of variance in first year and fourth year dental school GPAs, respectively. AA accounted for 31 and 22 percent of the variation in YR1GPA and YR4GPA, respectively. Whereas SCIGPA explained 47 and 45 percent of the variance in NBDEI and NBDEII, respectively, AA accounted for 28 percent of the variance in NBDEI and 20 percent of the variance in NBDEII.

Established by 1,875 subjects across five studies, the correlation of SCIGPA and NBDEI was the strongest among NBDEI predictors, explaining 47 percent of the variation in performance. However, OAGPA and AA explained 27 and 28 percent, according to 1,836 subjects in four studies, and 2,118 subjects in six studies, respectively.

Based on 1,476 subjects across four studies, SCIGPA was the best predictor of final cumulative dental school GPA (YR4GPA), explaining about 78 percent of variance. OAGPA achieved the next largest correlation with fourth year GPA, explaining 60 percent of the variation. This finding was derived from approximately 1,427 subjects in four studies. Nearly 1,500 subjects across 4 studies contributed to the association of DAT AA and fourth year GPA, which explained about 22 percent of variance.

Table 3-6: Corrected correlations, percent of variance explained and credibility intervals

Variable Pairing	Corrected Correlation	% variance explained	95 % Credibility Interval
AA:YR1GPA	0.561	31	0.489 - 0.632
AA:YR4GPA	0.474	22	0.419 - 0.529
AA:NBDEI	0.525	28	0.489 - 0.562
AA:NBDEII	0.443	20	0.402 - 0.485
PA:YR1GPA	0.322	10	0.237 - 0.408
PA:YR4GPA	0.249	6	0.187 - 0.312
PA:NBDEI	0.232	5	0.187 - 0.277
PA:NBDEII	0.258	7	0.212 - 0.304
TS:NBDEI	0.415	17	0.371 - 0.459
TS:NBDEII	0.389	15	0.341 - 0.437
OAGPA:YR1GPA	0.280	8	0.137 - 0.424
OAGPA:NBDEI	0.516	27	0.456 - 0.576
OAGPA:NBDEII	0.524	27	0.432 - 0.617
OAGPA:YR4GPA	0.774	60	0.719 - 0.828
SCIGPA:YR1GPA	0.564	32	0.439 - 0.688
SCIGPA:YR4GPA	0.837	70	0.785 - 0.888
SCIGPA:NBDEI	0.613	39	0.568 - 0.686
SCIGPA:NBDEII	0.669	38	0.540 - 0.687

Detailed calculations. In order to more clearly depict procedures and effects of each correction applied, formulas used and results for two variable pairings are presented.

Rosenthal-style calculations:

1. Individual *r*'s were transformed to z scores by formula:
   $Z_r = ½ \log e [1+r/1-r]$
2. Z scores were weighted by their sample size minus three and averaged
3. The average Z was transformed back to *r:*

First Author	r	z_r	Sample Size	z_r N_i
Behar-Hornstein	0.45	.485	209	98.848
Bergman	0.28	.288	249	70.770
DeCastro	0.325	.337	233	77.562
Fields	0.419	.459	451	134.836
Holmes	0.61	.709	566	399.123
Sandow	0.507	.559	410	227.384

The sum of z_r x N_i = 1083.181 was divided by the sum of N-3 (2115) and the result was a weighted-average z of .517 that was transformed back to an *r* of 0.472.

For SCIGPA:YR4GPA the same procedure was applied:

First Author	r	z_r	Sample Size	z_r N_i
Curtis	0.32	.332	49	15.256
Holmes	0.537	0.600	569	339.561
Fields	0.455	0.491	451	219.962
Sandow	0.425	0.454	410	184.688

The sum of z_r x N_i = 759.467 was divided by the sum of N-3 (1473) and the result was a weighted-average z of 0.514 that was transformed back to an *r* of 0.473.

Hunter& Schmidt-style calculations:

1. Mean *r* was calculated by formula**:** $\sum [N_i r_i] / \sum [N_i - 3]$

For Variable Pair AA:NBDEI, the data were as follows:

First Author	Correlation	Sample Size	$N_i r_i$
Behar-Hornstein	0.45	209	94.05
Bergman	0.28	249	69.72
DeCastro	0.325	233	75.725
Fields	0.419	451	188.969
Holmes	0.61	566	945.26
Sandow	0.507	410	207.87

The sum of $N_i r_i$ was 981.445 which as divided by the combined sample size (N-3=2,115) and the result was a weighted average *r*= 0.464.

For SCIGPA:YR4GPA, the data were:

First Author	Correlation	Sample Size	$N_i r_i$

Curtis	0.32	49	14.72
Fields	0.451	451	203.40
Holmes	0.61	566	345.26
Hermesch	0.429	361	154.87
Sandow	0.425	410	174.25

The sum of $\boldsymbol{N_i r_i}$ was 699.194 which as divided by the combined sample size (N-3=1,473) and the result was a weighted average $r = 0.475$.

2. Correction was applied for unreliability in the independent variable. The formula applied was:

$$rxy \ / \sqrt{(ryy_i)}$$

where ryy_i is the reliability of the dependent variable (NBDEI) for the incumbent population. For AA:NBDEI, this amounted to 0.464/(0.975) = 0.476.

For SCIGPA:DSYR4GPA, this amounted to 0.475/(0.866) = 0.548

3. Correction was applied for unreliability in the dependent variable. The formula applied was:

$$\boldsymbol{rxy \ / \sqrt{(rxx_i)}}$$

where rxx_i is the reliability of the independent variable in the incumbent population and is calculated using available data from rxx_a, the reliability of the independent variable in the applicant (unrestricted) group. By formula:

$$rxx_i = 1-[S^2x_a\,(1-rxx_a)]/\ S^2x_i$$

Substituting into the formula, for AA:NBDEI, it was found that

$$rxx_i = 1-[4.41\,(0.05)]/\ 3.71 = 0.941$$

Substituting into the formula, for SCIGPA:YR4GPA, it was found that

$$rxx_i = 1-[0.25\,(0.25)]/\ 0.149 = 0.580$$

Therefore, the correlations already corrected for reliability in the dependent variable were further corrected by dividing by the square root of rxx_i:

AA:NBDEI r_c=0.488/0.970= 0.491
SCIGPA:YR4GPA r_c=0.488/0.761= 0.720

4. Corrections were applied for range restriction.

First the mean value of u_T was calculated:

$$u^2_T = [u^2_X\text{-}(1\text{-} rxx_a)/ rxx_a$$

where $u^2_X = sx^2/Sx^2$, that is the ratio of the observed variance in the restricted sample to the observed variance in the unrestricted sample. Substituting AA:NBDEI data into the formula,

$$u^2_T = [0.841\text{-}(.05)/ 0.95= 0.833$$

$$u_{T=} \sqrt{u^2_T} = 0.913$$

Substituting SCIGPA:DSYR4GPA data into the formula,

$$u^2_T = [0.595\text{-}(.25)/ 0.75= 0.460$$
$$u_{T=} \sqrt{u^2_T} = 0.678$$

These values were used to apply the correction for range restriction:

$$PTP_a = U_T\ PTP_i/[1+ U^2_T P^2TP_i – P^2TP_i]^{1/2}$$

(Hunter, Schmidt & Le, 2006)

Substituting into the formula, and where $U_T = 1/u_T$

AA:NBDEI PTP_a = (1.095)(0.491)/[1+0.833(0.241)-(0.241)]$^{1/2}$ = 0.525
SCIGPA:YR4GPA PTP_a = (1.47)(0.720)/[1+2.372(0.518)-(0.518)]$^{1/2}$ = 0.837

Comparison of Uncorrected and corrected Correlations. Corrections for measurement error and indirect range restriction yielded correlations that were between 10 and 131 percent larger than before adjustments.

Potential for publication bias

Table 3-7 presents bivariate relationships between uncorrected correlations and sample size found in primary studies. It shows strong positive relationships between sample size and correlation. The same correlations could not be computed using corrected correlations because artifact distribution methods were applied.

Table 3-7
Number of studies, relationship between correlation and sample size in primary studies

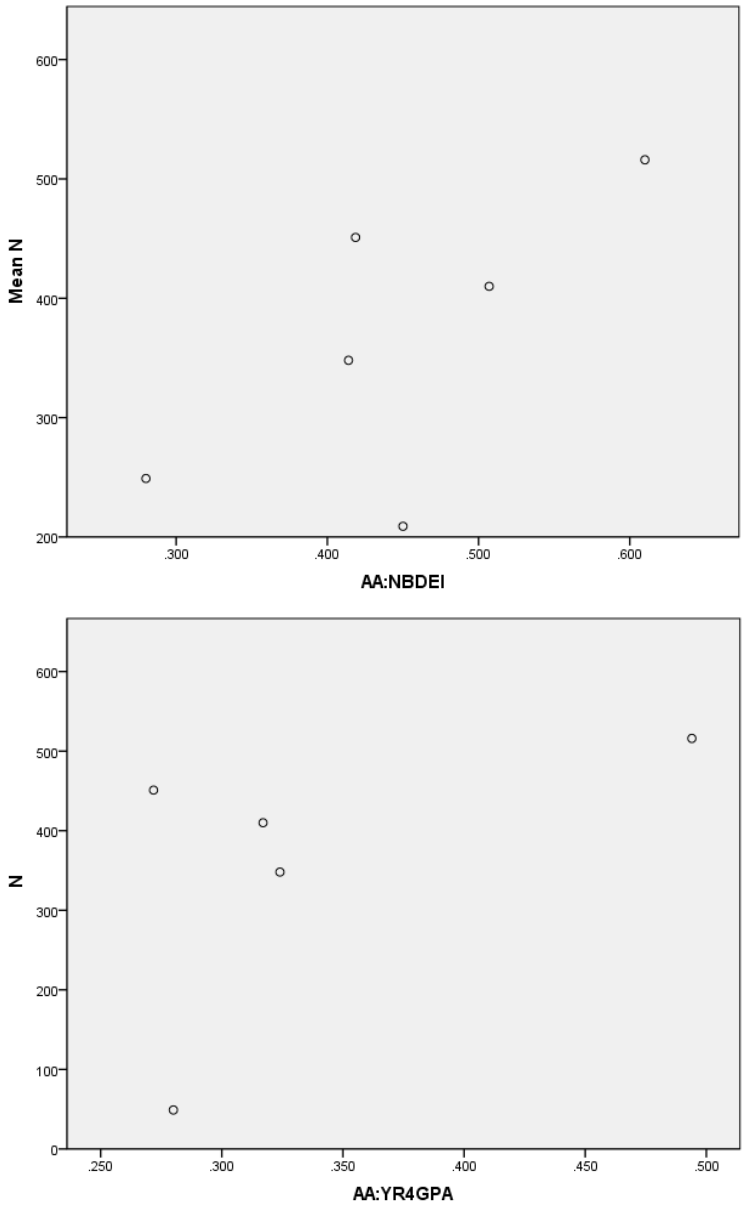
Number of Studies	Correlation Pair	Correlation between Uncorrected correlations and N
3	AA:YR1GPA	0.727
4	AA:YR4GPA	0.540
6	AA:NBDEI	0.647
6	AA:NBDEII	0.805
3	PA:YR1GPA	0.999
3	PA:YR4GPA	0.790
7	PA:NBDEI	0.820
6	PA:NBDEII	0.792
4	TS:NBDEI	0.792
4	TS:NBDEII	0.743
3	OAGPA:YR1GPA	0.980
4	OAGPA:YR4GPA	0.958
5	OAGPA:NBDEI	0.671
4	OAGPA:NBDEII	0.833
4	SCIGPA:YR1GPA	0.308
4	SCIGPA:YR4GPA	0.940
4	SCIGPA:NBDEI	0.736
5	SCIGPA:NBDEII	0.987

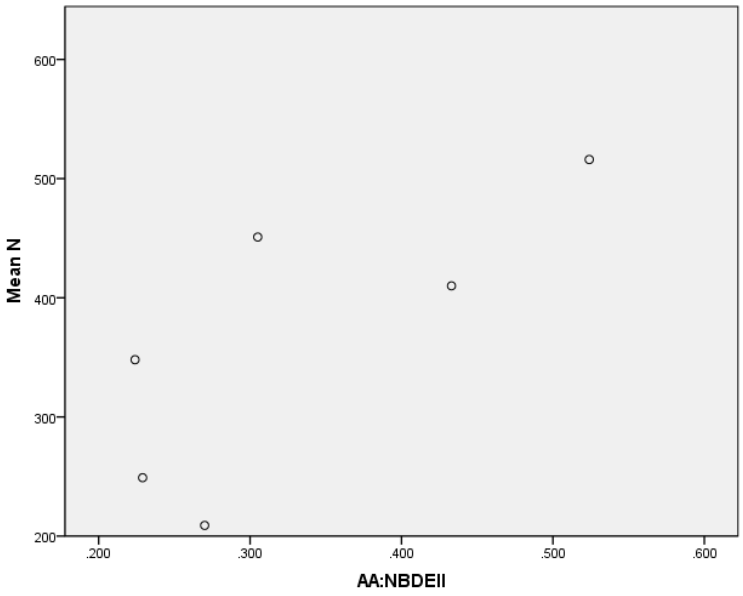
Figure 3-2 presents a plot of uncorrected correlations on the vertical axis as a function of sample size on the horizontal axis. Large studies appear toward the top of the

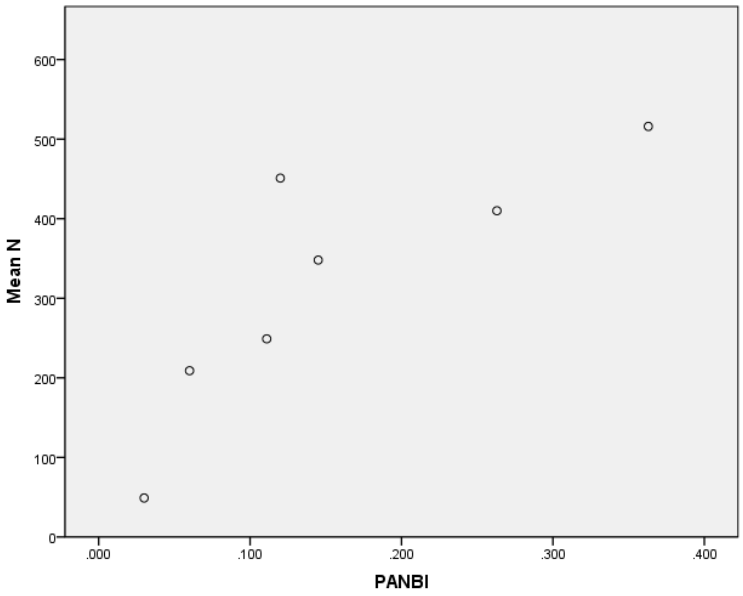
graph, and larger correlations to the left. Smaller studies appear toward the bottom of the graph. If a "small study effect" were present, smaller correlations would be associated with larger effect sizes, and values would be clustered in the upper left of the graph, which does not seem to be the case. Figure 3-2 also graphically depicts the patterns identified in table 3-7.

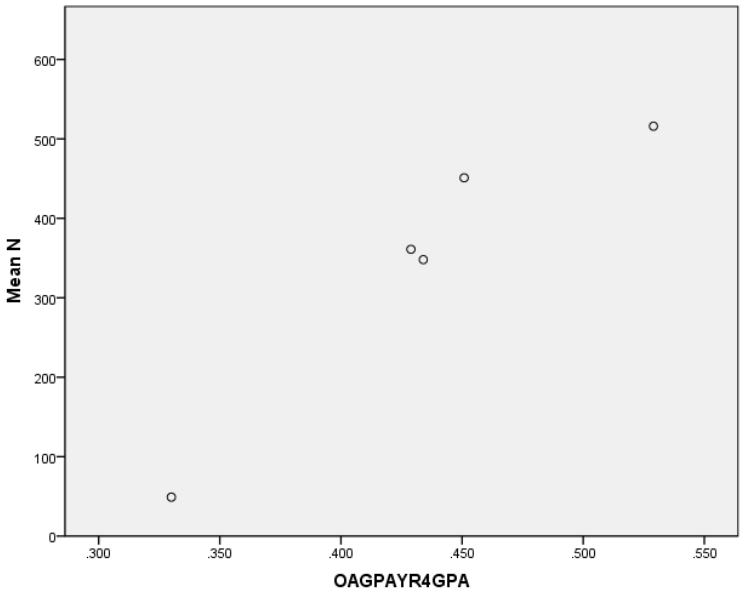
Figure 3-2

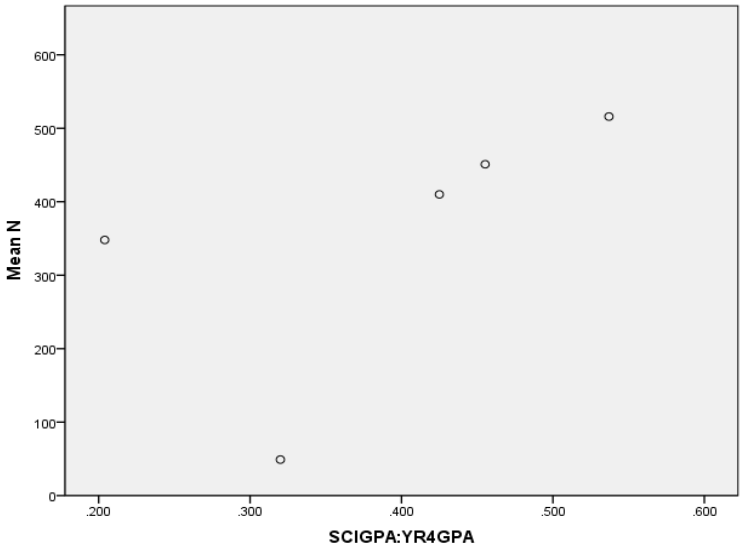
Graphs of sample size to uncorrected correlations in primary studies

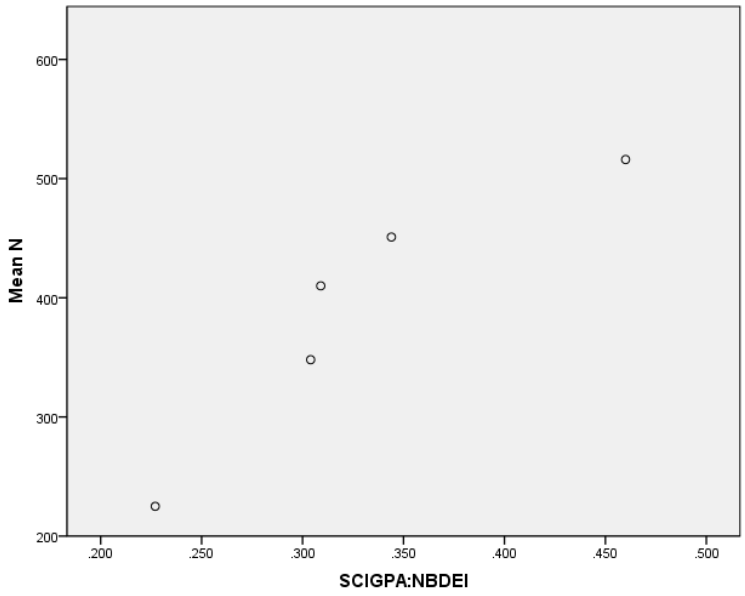


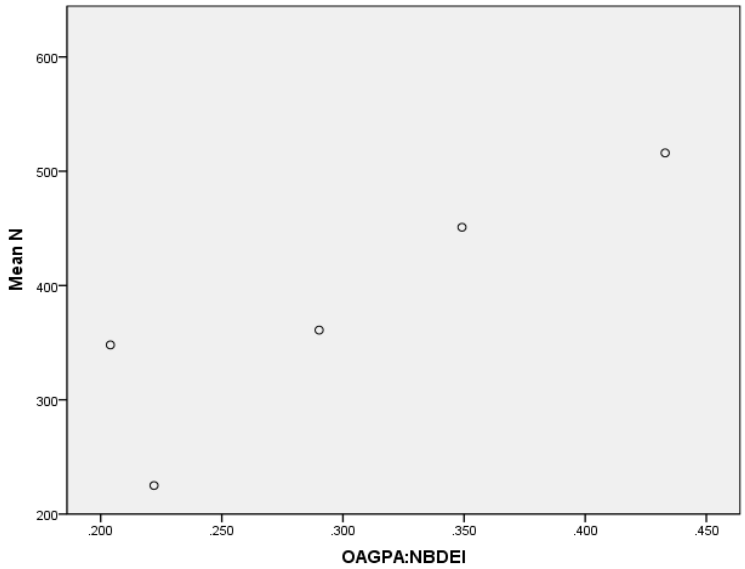

Mean N
600
500
400
300
200
.200
.300
.400
.500
.600
AA:NBDEII


Mean N
600
500
400
300
200
100
0
.000
.100
.200
.300
.400
PANBI


Mean N
600
500
400
300
200
100
0
.300
.350
.400
.450
.500
.550
OAGPAYR4GPA


Mean N
600
500
400
300
200
100
0
.200
.300
.400
.500
.600
SCIGPA:YR4GPA



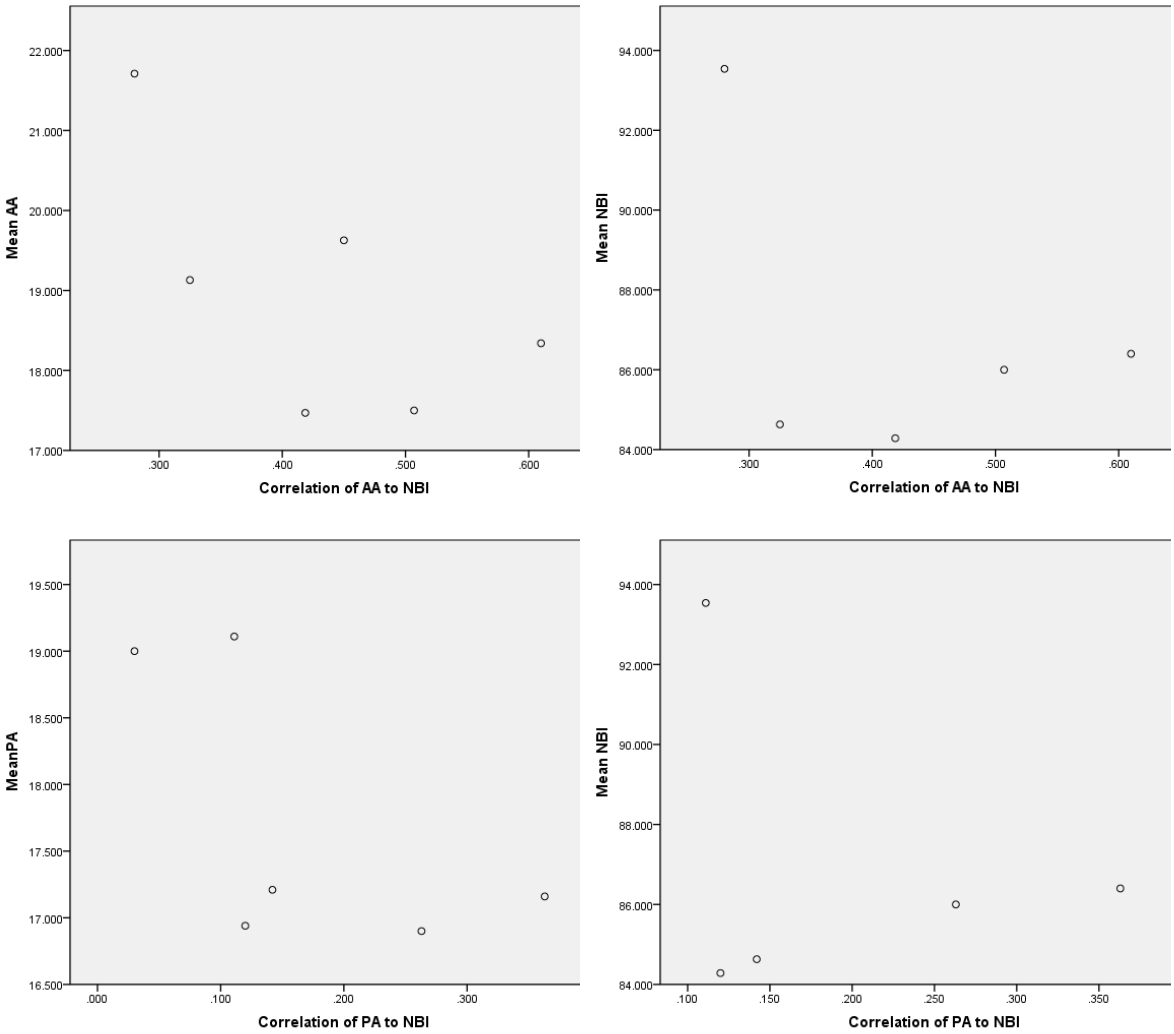


The search for moderators

Even though population variance was estimated as being quite small, planned procedures to account for residual variance were still conducted. Figure 3-3 presents a visual display of the correlations found in primary studies against DAT AA and PA scores, NBDEI score and OAGPA and YR4GPA, as a means of examining whether the correlation tends to change over the range of scores or GPAs. Although the highest scores

tend to be associated with lowest correlations, the lowest scores are not consistently associated with the highest correlations.

Figure 3-3 *Graphs of scores against uncorrected correlations*



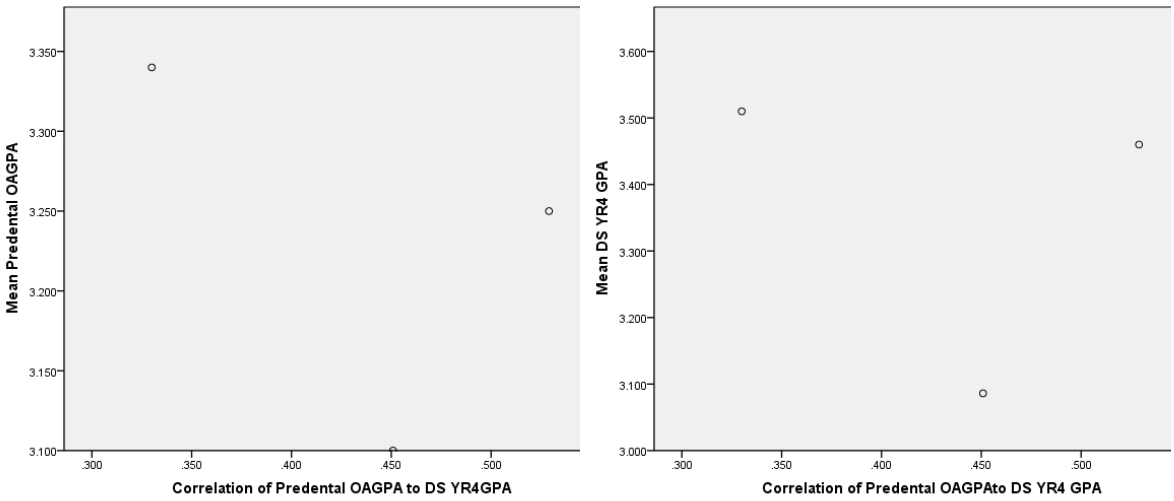
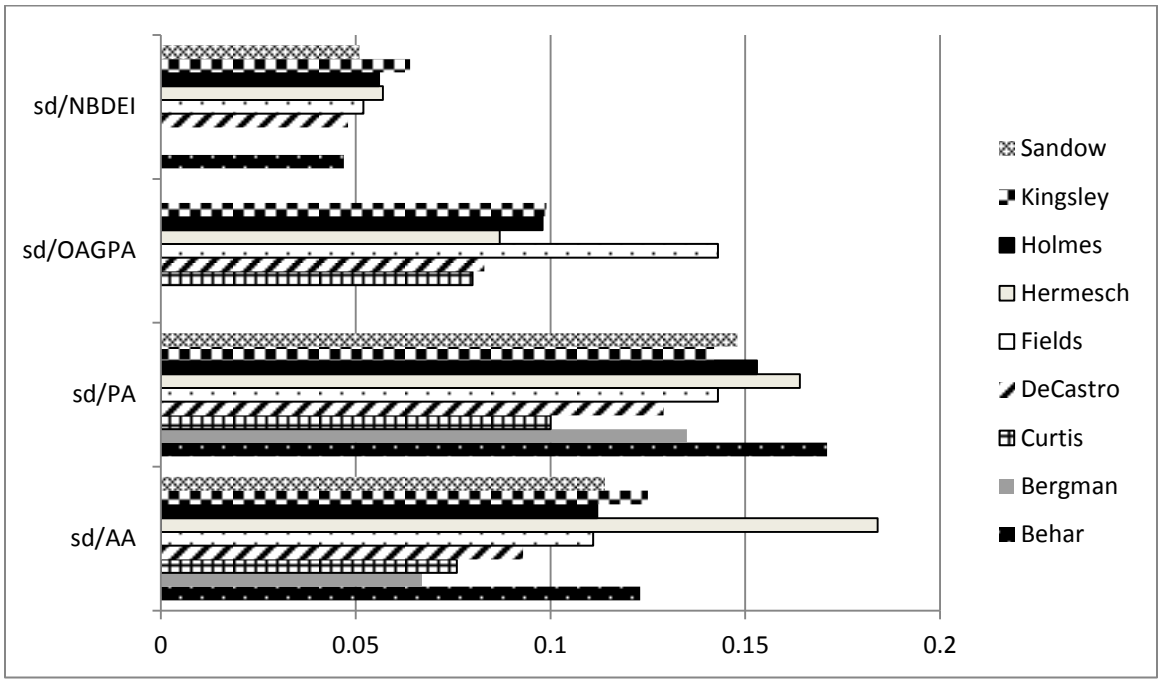


Figure 3-4 presents the proportion of variability represented by the ratio of standard deviation to score.



There were not clear and consistent relationships found among the relationship between the ratios of standard deviations to score. Among the criterion variables, there were only very slight differences in variability to score.

Table 3-8 presents school data concerning class size, selectivity, curricula type and prerequisite courses. These were several of many school-level variables examined for purposes of exploring whether these factors were associated with correlation magnitude. Authors have suggested that these factors may influence student performance. Clear patterns were not found.

Table 3-8

Entering class size, percent of applicants enrolled, curriculum and admissions prerequisites

School	Entering class size	% of Applicants Enrolled	Curriculum Type	Admissions Prerequisites besides 1 year of General Biology with lab 1 year of General Chemistry with lab 1 year of Organic Chemistry with lab 1 year of General Physics with lab
UNevada-LV	78	2.93	Traditional	1 semester of Biochemistry 1 year of English
Harvard	35	3.70	PBL	1 semester of Biochemistry 1 year of English 1 year of Calculus
UMDNJ	88	4.05	Traditional	1 year of English
UCSF	83	4.80	"Thematic Streams"	1 semester of Biochemistry 1 year of English Composition 1 year of Calculus 1 semester of Psychology 2 years of social sciences, humanities of foreign language,
UFlorida	82	5.88	"Thematic Streams"	1 semester of Biochemistry 1 year of English Grammar/Composition 1 year of Calculus 1 semester of Microbiology 1 semester of Molecular Biology 1 semester of Psychology
UIowa	77	7.74	Traditional	1 year of English
OhioState	101	9.42	Traditional	[only 1 semester of General Physics with lab required] 1 year of English 1 semester Biochemistry 1 semester Anatomy 1 semester Physiology
UTexas-SA	93	15.27	Traditional	1 year of English 1 semester Biochemistry 1 semester statistics

Chapter 4: Discussion

Although assessing the validity of predictors of dental school performance has received scholarly attention over the years, when looked at collectively in a narrative review Ranney, et. al., (2005), identified a wide range of correlations, making interpretation difficult. Available methods to correct for artifacts of measurement had not been applied, nor had there been any empirical investigations to identify factors that may influence the strength of these relationships.

Combined correlations

The first purpose of this dissertation was to synthesize research results to produce mean correlations of common predictor and outcome variables.

Comparison to other graduate level studies. Referring back to the brief review of meta-analyses of predictors of performance in other disciplines of graduate education presented in Table 1-1, the expectation was that the corrected correlations would be similar in magnitude to those found in other meta-analyses. However, it should be noted that the Donnon et. al., 2007 study corrected for restriction of range, but no adjustments were made for unreliability of measures, thus correlations for MCAT were not as disattenuated. In the meta-analyses of medical school performance, MCAT correlations appear to follow consistent patterns to those found for DAT scores. Corrected correlations for MCAT total and USMLE I scores reported by Donnon et. al., (2007) were higher, but still comparable (0.66) to corrected correlations for AA and NBDE I using Hunter & Schmidt methods (0.53). Like MCAT scores, DAT scores generally seemed to explicate less variability over time, or rather to explain more of preclinical than clinical performance, with the correlation for MCAT and medical school preclinical

and clinical grades reported by Donnon et. al at 0.43 and 0.39, respectively. For DAT, the finding was 0.70 and 0.55 for the corrected correlations of AA with first and final cumulative GPAs. Akin to MCAT overall's correlation with USMLE Step 3, (0.48), DAT AA explained less in NBDEII (fourth year) scores (0.45). The findings of Kreiter & Kreiter, 2007, included MCAT and premedical GPA as predictors, but corrected for reliability and not range restriction. The findings again displayed similar patterns with MCAT and premedical grades explaining 0.61 of first and second year exam scores. At the same time, corrected correlations reported for more general tests, like GMAT total with first year GPA reported by Kunzel et al, 2007, although corrected for reliability and restriction of range, were smaller (0.47) than those found involving DAT. Kunzel, et al., 2001 did not report a GRE total, but reported GRE subject test correlations and undergraduate GPA correlations with first year GPA and comprehensive exam scores. These too were comparable, albeit smaller than those found with DAT and predental grades with dental school performance.

Although it is at first tempting to interpret this as evidence that testing abilities and knowledge more closely related to the performance being predicted is superior to testing based on more general abilities, this is not proven, and alternative explanations are plausible. Differences among the MCAT/DAT and GRE/GMAT correlations were much smaller before corrections were applied. While reliabilities for GMAT, GMAT, DAT and MCAT and associated grades are nearly identical, it may be that the restriction of range is much greater among the dental and medical school populations than that found among the general population of graduate applicants, making adjustments for restriction of

range, and consequently, corrected correlations, correspondingly larger. Whether these adjustments reflect real differences or artifacts of method is unclear.

Uncorrected combined mean correlations. Consistent with primary studies, findings using the Rosenthal method showed AA scores were better at predicting performance measured sooner (first year GPAs, NBDEI) as compared to variables measured on a longer term basis (final GPAs and NBDEII). Although DAT AA, which includes TS and more general ability scales of RC and QR, had a higher correlation overall with NBDEI than TS alone, the use of AA (which adds the effects of RC and QR) increased explained variance by less than two percent for NBDEII.

While thought to be a measure of cumulative achievement in the sciences as well as reading and quantitative abilities, AA may reflect at least some written test-taking ability, and that as the curriculum advances from written to clinical competency tests, written test-taking (along with reading and quantitative) ability becomes less relevant.

If one looks at the AA composite score in isolation, it is possible for very high scores in the RC and QR components to increase AA to the point that correspondingly lower performance on the TS scales is masked. It is not clear how often this occurs. While it would seem obvious that the ability to absorb new scientific material through reading should greatly influence performance, this may not hold true in all situations. If high reading comprehension ability did not result in demonstrated achievement in the sciences at the undergraduate level as evidenced by TS score, it may be unlikely that reading skills alone (without a matching record of achievement in the sciences) are enough to carry the day at the professional level. While this seems to contradict findings using regression analysis showing RC was the strongest predictor of some NBDEI

subscales (Bergman, et. al., 2006; DeBall, et. al., 2002), there were not enough primary studies reporting correlations involving RC to explore its usefulness in relation to TS.

Although first year grades are of interest, prediction of student success across the full curriculum (including preclinical and clinical years) is of paramount concern, second only to success in practice, which is beyond the scope of this study. Despite all the fallibilities of GPAs –even before adjustments for reliability-- predental GPAs were found to be superior over DAT AA in predicting fourth year GPA.

Disattenuated correlations. As expected, disattenuated correlations were larger than uncorrected correlations, but relationships among them remained relatively similar. Adjustments due to unreliability were greatest for correlations involving grades as either predictors or outcomes, and this caused those correlations to increase greatly. Among the corrected correlations, SCIGPA was consistently the strongest predictor. For example, SCIGPA and OAGPA explained 70 and 60 percent of variance in fourth year GPA, respectively. SCIGPA explained about 32 percent of the variance in first year GPA, followed by AA which explained 31 percent. Although SCIGPA remained the best predictor of NBDEI, OAGPA explained about the same amount of variance in NBDEI as AA did. Adjustments for indirect restriction of range were greater for independent variables than for dependent variables, as the selection effects were substantially greater than attrition effects. In addition, such adjustments were greater for grades since there were larger differences in variability between incumbents and applicants.

Estimates of reliability of correlations and amount of variance attributed to sampling error (which have an inverse relationship) add context to the findings, and suggest further caution in interpreting disattenuated correlations. Correlations that are

either very unreliable or highly influenced by sampling error are suggestive of less trust being placed in some predictors – the relatively larger credibility intervals further enhance such interpretations.

Reflecting back on the supposition in the introduction, that analysis of dental school performance presented a hybrid of prediction of assimilation and application of knowledge, once again availability of data is a factor. Although theoretically first year GPAs in many dental schools assess assimilation of knowledge, since grades are mostly based on basic science performance, cumulative final dental school GPAs incorporate these grades with those of second thru fourth years, rendering the criterion impure with regard to assessment and application of knowledge. Availability in primary research of predictors' correlations with yearly GPAs across the dental school curriculum, or even further differentiated by basic science, technique and clinical course averages, would permit more detailed analysis of this idea. For now, it is only apparent that corrected SCIGPAs appear to be best among the predictors at forecasting YR4GPAs.

Comparison of uncorrected and corrected correlations. In all variable pairings, the mean correlation was more than two (corrected) standard deviations larger than zero, reasonably permitting the conclusion that the relationships are always positive (Hunter & Schmidt, 2004). As noted, corrected correlations were larger and those with grades in the variable pairing were much larger due to lower reliability of grades and greater differences in variability in the incumbent and applicant populations. The practical use of knowing how well DAT scores or grades would predict performance if they were perfectly measured is unclear. There may be cases where an admissions officer has established the grading of certain feeder schools to be highly reliable, which might make

the information more practical. However, Hunter & Schmidt (2004) suggest as researchers the goal is to understand the relationship among constructs, and that may be of some benefit.

Identification of potential moderator variables

The second purpose of this dissertation was to attempt to identify school-level variables that affect the magnitude of the relationships under study. Once variance related to sampling error was removed, the remaining variance was nominal, suggesting that much of the differences between studies was attributable to sampling error. Since corrections for unreliability with KR-20 estimates do not address transient error, it is possible that at least some remaining variance in correlations involving DAT and NBDE variables is due to incomplete corrections for reliability. Furthermore, remaining variance may be a function of the small number of studies found (because sampling error was divided among so few studies).

Still, it is possible that the unreliability in dental school grades encompasses some of the characteristics of school-level moderator variables. That is, when one takes into account the measurement error inherent in dental school GPAs, one seems to control for much of the school-level variation, since little unexplained variation remained in the prediction of fourth year grades by predental grades.

Analysis of the ratio of standard deviation to score shown in Figure 3-4 did not consistently produce clear groupings. The ‘selectivity ratio’ or percentage of applicants to enrolled students in table 3-8 did not help identify selectivity as a possible moderator, but it is not definitively excluded. There could be more subtle differences in force that are

incalculable from available data: it was not possible to calculate the ratio of offered positions to enrolled students, a number that might more accurately assess selectivity.

There did not seem to be a corresponding increase in prerequisites for schools showing higher selectivity, nor distinctions among the correlations for schools with fewer or more prerequisites. However, it should be noted that some schools increase prerequisites (such as Microbiology) in order to reduce curriculum time devoted to that topic, so addition of a prerequisite may not necessarily translate into stronger preparation or performance. Furthermore, although some schools require students to take more courses before arriving than others, other schools list such courses as "recommended." Not knowing how much of the student body completed these courses, it is impossible to determine if these resulted in a better prepared entering class.

Comparisons based on curriculum type may have suffered from the same lack of complete or accurate information. The explanation for differences not being found could include either there is no relationship or, the lack of uniformity in reporting curricula types. Other potential moderators explored included type of funding (private or public); whether students were required to pass a qualifying exam in order to challenge NBDEs; and, whether schools accepted students with three years of college or required a Bachelor's degree. Several schools offer seven year programs, and Hermesch et. al., (2005) found significant differences in performance between traditionally admitted and early acceptance students. Yet, there did not seem to be differences between schools that offered such programs and those that didn't. The two schools that reported figures for this reported very low participation (for example, 3 out of 88 students). Consequently differences may be lessened. Even though schools may officially permit such applicants,

in practice they may rarely if ever accept students without Bachelor degrees. In summary, the exploration of potential moderator variables did not produce definitive results. It is hoped that additional primary studies and availability of more data in the future might lead to a more fruitful pursuit.

Measurement issues

The third purpose of this work was to identify issues that inhibit data synthesis and develop reporting conventions that would facilitate future efforts at synthesis. As noted earlier, availability of grades for each year of curriculum separately might permit better assessment of which predictors are better at forecasting what aspects or at least years of curriculum. Similarly, research reporting correlations with basic science, psychomotor and clinical grades would permit more refined comparisons.

Many individual characteristics that could influence strength of the relationship between predictors and outcomes have been identified in prior research. It would be helpful to clearly define proportionate representation and correlations of subgroups (such as repeat test-takers, students who've completed graduate work, gender, minority status and admissions program) in populations studied in order for group differences to be accounted for in interpreting results across populations. Similarly, due to lack of uniformity in how failing grades are reported (or not reported) by institutions that applicants attended, use of GPAs as predictors or criterion will remain problematic, making it more critical for researchers to be painstaking in describing methods. It would be helpful if the exact source of data were reported as well, such as first time test results, as well as whether plusses or minuses or graduate work were included in the GPAs. Likewise, it would be useful when reporting NBDE correlations and results to know if

students were required to pass a qualifying examination, as these results also could be expected to be systematically different.

The fact that KR-20, the reliability estimate provided for both DAT and NBDE scores does not account for all possible error types (Hunter & Schmidt, 2004), may mean that reliability coefficients reported by ADA were larger than accurate by the amount of unaccounted for transient-type errors. Consequently, corrections for reliability for DAT and NBDE scores were probably smaller than they should have been, and therefore corrected correlations involving these variables may be biased downward.

The NBDEs are expected to become scored on a pass/fail basis beginning in 2012, and future researchers will not have easy access to continuous scores like the present standardized scores. Moving to comparison of continuous predictor scores against dichotomous (pass/fail) criterion may reduce power and further limit published articles using NBDEs criterion. On the other hand, there may be less motivation to withhold publication from schools previously not inclined to report NBDE averages.

More importantly, however, if data from individual studies were retained and de-identified data shared amongst researchers—perhaps in a central repository with the American Dental Education Association (ADEA) or ADA, such analyses could be performed much more readily and accurately, regardless of whether the primary researcher has retired or left the institution. There were several publications that appeared to have collected data but not reported correlations of interest, but there were problems locating the data. Last, if ADA would provide researchers with dis-identified school-level data, research in this area could advance.

Limitations

Potential biases in retrieved literature. It must be acknowledged that the number of studies that comprise this meta-analysis is quite small and there are potential differences between the identified and unidentified studies, as well as lack of stability in the findings. Although correlations need only be quite weak (0.16) to produce significant findings when working with populations that are reasonably large (n=100), it is possible that (likely smaller) studies with no correlations or those that found insignificant correlations were not published. However, data displayed table 3-7 and in figure 3-2 suggest that studies with larger sample sizes tended to be associated with larger correlations. Consequently, these data seem to contraindicate the "small study effect" theory, which suggests that smaller studies are more likely to be published if they have larger than average effects, since they would be more likely to meet the criterion for statistical significance. Efforts were made to locate unpublished studies, but this work was nonetheless limited by the small number of primary studies and effects of publication bias.

In addition, based on the finding that only studies reporting scores equal or higher than national median scores were located, it is possible that a distinct type of publication bias is in play: studies that may have presented predictor or criterion values lower than the national median for year of publication were not found. This could reflect a self-censoring effect of schools not wanting to "publicize" lower performance on national licensing exams or even incoming GPAs or DAT scores, as such statistics are often (correctly or not) thought to reflect school quality. Consequently data collected may represent only the upper half of the distribution, and patterns found may not be the same

as those operating on the whole, and should not be extrapolated as representing the whole.

Lack of data. This work suffered from lack of available information in several ways, such as reliability estimates of dental school grades, estimates of reliability in individual primary studies and from all schools nationally. The lack of data in primary studies resulted in adjustments being made via artifact distribution methods which are slightly less accurate than corrections made with individualized data. For instance, corrections for restriction of range were applied uniformly based on the mean data. The restriction may actually have been much greater at more selective schools, and lower at less selective schools, yet it is applied unvaryingly, thus over- and under-correcting data at the extremes.

Although the ADA collects and reports data from all dental schools concerning correlations between DAT and predental GPA and performance during the first two years of dental school, this is done without revealing the identity of schools or their sample size for listed data. The ADA did not feel it was at liberty to release data such as sample sizes of the various reported correlations that would have permitted further exploration of patterns, thus thwarting attempts to identify moderators.

Correlations in which GPAs were either in the independent or dependent variables (or both) tended to increase more from adjustments due to unreliability, as the standardized tests (DAT and NBDE) reported reliabilities that were relatively high. Adjustments involving dental school GPAs in the dependent variable rested on the untested premise that reliability for dental school grades is similar to that of medical school and college grades. To the degree that dental school grades might be actually

substantially more reliable than estimated, these adjustments may have been biased upwardly.

This dissertation was restricted by limitations inherent in correlational studies, correlational meta-analyses as well as in this particular one. Inasmuch as the predictor variables are highly intercorrelated, regression and hierarchical methods would be more useful in modeling the predictive relationships, albeit these more sophisticated methods do not lend themselves as readily to later quantitative synthesis. As correlational data the findings are specific to these samples and times, and can't be generalized, regardless of whether fixed or random models were applied.

The findings relating to sampling error were inconsistent. For instance, a high (69) percentage of variance in the correlations of overall predental GPA and first year dental school GPA attributed to sampling error combined with the low percentage of sampling error found in correlations involving relationships with fourth year GPA could be interpreted as suggesting all applicants are drawn from the same population and later performance was diluted by differences among schools. However, the variation due to sampling error in the correlation of predental science GPA and first year dental school GPA was only 19 percent. Since the two predictors (OAGPA and SCIGPA) were so closely related one would expect more similar results. Sampling error would generally be expected to be least among correlations based on the largest number of studies and most among the correlations based on fewer studies, but this was not consistently found. It may simply be the case that the sample of studies is too small to produce stability in results.

The subjects represented in the included studies did not represent a sizable portion of the nearly 14,000 applicants (times 20 years) to dental schools and were limited in

number. Due to these limitations, any extrapolation of results to other populations is risky.

Time confounds. Furthermore, combining results that change over time (such as the upward trend in AA scores noted earlier) from studies based on data from different years most likely led to some confounding with time. Although this is unavoidable when conducting a meta-analysis across approximately two decades of primary research, the process added time as an artifact of measurement that was not as substantial in the original studies.

Future research

Strong relationships were found between sample size and correlations, a finding that seems worthy of further pursuit. The finding that published works in this area may represent the upper end of the distribution of scores rather than the full spectrum implies the need for more primary research representing the lower distribution of scores and subsequent synthesis. Additional primary research could result in sufficient cases to perform a productive cluster analysis, a necessary step toward identifying moderator variables.

The lack of stability in correlations involving underrepresented minority groups on a national level (Kramer 1999) urgently needs to be confirmed both locally and nationally. Until such time as this information is published or released, how much confidence to place in the predictive utility of DAT scores for underrepresented minority students is indeterminable.

Summary and conclusions

If nothing else, results suggest DAT and predental grades are likely much stronger predictors of dental school performance than previously reported overall, but results vary among schools due in part to sampling differences, and in part for other reasons not yet fully understood. As hypothesized, uncorrected correlations of AA with various measures of performance ranged from 0.37-0.48, or medium to high medium, and after corrections, correlations grew to 0.45 to 0.70 or medium to high. The uncorrected correlations of various predental GPAs with dental school performance indicators ranged from 0.15 to 0.44, low medium to medium, and after adjustments these ranged from 0.28 to 0.84. With the exception of a surprisingly low correlation found between Overall predental GPA and first year dental school GPA (0.28), all corrected correlations involving predental grades were in the medium high (0.41) to high (0.84) range. It should be noted that the number of studies the overall predental GPA to first year dental school GPA was quite low, and accuracy of results suffered for it, as reflected in its high sampling error.

Fortunately, admissions officers don't have to rely on only one predictor, and in fact, decisions that consider GPAs and DAT scores are much sounder than decisions that use one alone. If results were more reliable and could be applied more widely, it could be said that they held the following implications: when results between DAT scores and GPAs are disparate, if one were more interested in longer term criterion and had reason to believe that the GPAs in question were highly reliable, one might give slightly more weight to GPAs. If for some reason one preferred to increase first year GPAs and NBDE I, one might give slightly more weight to AA when scores are inconsistent with GPA.

This dissertation intended to clarify understanding and estimation of correlations between identified predictors of dental school performance and criterion and then to

identify potential moderators to those relationships. From nine primary studies, mean correlations and confidence intervals were calculated, and after corrections for range restriction and measurement error, corrected correlations and credibility intervals were obtained that can be used to enhance interpretation of results found at the local level. Although multiple limitations are present as defined above, the results included in this investigation may provide added perspective using statistical tools that had not previously been applied to evaluation of correlations among these variables. It is hoped that future research will overcome these confines and provide more accurate true correlation estimation and definitive identification of moderator variables.

Table 3-1

Means and standard deviations of dental school admission criteria and dental school performance indicators

1st Author, Year		Behar-Hornstein 2011	Bergman, 2008	Curtis, 2007	DeCastro, 2010	Fields, 2003	Hermesh 2005	Holmes, 2008	Kingsley, 2007	Sandow, 2002
School Sample size	Variable	UFlorida 209	Harvard 249	UCSF 49 normally tracking	UMDNJ 351	OhioState 451	UTexas 361	UIowa 566	UNevada 210	UFlorida 410
Mean±sd of	DAT AA	19.67±2.2	21.71 ± 1.45	21 ± 1.6	19.13±1.79	17.47±1.93	17.82±1.49	18.34 ± 2.1	18.00 ± 2.50	17.5 ± 2.0
Admissions	DAT PAT	17.99±2.19	19.11 ± 2.58	19 ± 1.9	17.21±2.22	16.94±2.42	17.44±2.85	17.16 ± 2.6	18.50 ± 2.63	16.9 ± 2.5
Criteria:	DAT TS	19.28±2.69	21.57±1.59		19.18±1.84	17.07±2.09	17.41±2.16	18.13±2.18		
	OAGPA		3.60 ± 0.26	3.51 ± 0.28	3.47±0.29	3.09±0.44	3.35±0.29	3.46 ± 0.34	3.34 ± 0.33	
	SCIGPA		3.63 ± 0.23	3.46 ± 0.33	3.41±0.34	3.19±0.38		3.47 ± 0.41	3.22 ± 0.39	3.0 ± 0.40
Mean±sd of	YR1GPA			3.43 ± 0.28	3.11±0.48				3.39 ± 0.47	3.2 ± 0.50
Performance	YR4GPA			3.34 ± 0.27			3.34±1.45	3.25 ± 0.39		3.2 ± 0.3
Measures	NBDEI	87.74±5.80	93.54 ± 3.63		84.63±4.09	84.37±4.40	85.60±4.40	86.4 ± 4.8	84.76 ± 5.42	86.0 ± 4.4
	NBDEII	82.945±5.62	84.26 ± 4.30		80.62±4.07	81.23±4.77	82.26±4.66	83.9 ± 4.6		81.8 ± 4.2

Note. For the Curtis et. al., 2007 study, results of the 45 students whose performance did not follow usual patterns were excluded.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Alon, S., & Tienda, M. (2005). Assessing the "mismatch" hypothesis: differences in college graduation rates by institutional selectivity. *Sociology of Education, 78(4),* 294-315.

American Dental Association, Department of Testing Services, 2011. *Dental admissions testing program report 3, user's manual, 2009.* Chicago: IL: Author. http://www.ada.org/prof/ed/testing/dat/dat_users_manual.pdf accessed March 17, 2011.

American Dental Association, Department of Testing Services, 2006. *Dental admissions testing program report 3, user's manual, 2005.* Chicago: IL: Author. http://www.ada.org/prof/ed/testing/dat/dat_users_manual.pdf accessed November 15, 2008.

American Dental Association, Department of Testing Services, 2008. *Dental admissions test candidate guide.* Chicago: IL: Author. http://www.ada.org/prof/ed/testing/nbde01/nbde01_candidate_guide_2008.pdf

American Dental Association, Department of Testing Services, 2007. *Dental admissions testing program report 1 2007.* Chicago: IL: Author.

American Dental Association, Department of Testing Services, 2008. *Dental admissions testing program report 1 2008.* Chicago: IL: Author.

American Dental Education Association (2010). *2010 ADEA Official Guide to Dental Schools*. Washington, DC: ADEA.

Barritt, L.S. (2006). Class attendance and gender effects on undergraduate students' achievement in a social studies course in Botswana. Essays in Education, 18, 1-11.

*Behar-Hornstein, L.S., Garvan, C.W., Bowman, B.J., Bulosan, Hancock, S., Johnson, M., Mutlu, B., (2011). Cognitive and learning styles as predictors of success on National Board Dental Examination. *Journal of American Dental Education Association, 75(4),* 534-543.

*Bergman, A.V., Susarla, S.M., Howell, T.H., & Karimbux, NY (2006). Dental Admission Test scores and performance on NBDE Part I, revised. *Journal of Dental Education, 70(3),* 258-262.

Bureau of Labor Statistics (2009). http://www.bls.gov/oes/current/oes_nat.htm#b00-0000 and http://data.bls.gov/PDQ/servlet/SurveyOutputServlet;jsessionid=f030230b744aN$3F$3F$ accessed on February 7, 2009).

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

Cooper, H. (2010). *Research synthesis and meta-analysis a step-by-step approach, 4th edition*. Thousand Oaks, CA: Sage.

*Curtis, D.A., Lind, S.L., Plesh, O., & Finzen, F., (2007). Correlation of admissions criteria with academic performance in dental students. *Journal of Dental Education 71(10),* 1314-1321.

De Ball, S., Sullivan, K., Horine, J., Duncan, W.K., & Replogle, W. (2002). The relationship of performance on the Dental Admission Test and Performance on Part I of the National Board of Dental Examinations. *Journal of Dental Education 66(4),* 478-484.

*DeCastro, J.E. (2010). Structural equation model of dental school predictors and performance. Unpublished manuscript, University of Medicine and Dentistry of New Jersey.

Divgi, D.R. (2005). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23(4),* 283-298.

Donnon, T.D., Paolucci, E.O., & Violato, C. (2007). The predictive validity of the MCAT for medical school performance and licensing board examinations: A meta-analysis of the published research. *Academic Medicine, 82(1),* 100-106.

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

*Fields, H.W., Fields, A.M., Beck, F.M.(2003). The impact of gender on high stakes dental evaluations. *Journal of Dental Education: 67(6):* 654-660.

Geiser, S., & Studley, R. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment, 8(1*),1-27.

Glass, G.V. (1980). Summarizing effect sizes. In R. Rosenthal (Ed.). *New directions for methodology of social and behavioral science: Quantitative assessment of research domains.* San Francisco: Jossey Bass.

Hall, J.A., Tickle-Degnen, L., Rosenthal, & R., Mosteller, F. (1994). Hypotheses and Problems in Research Synthesis. In H. Cooper & L.V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 17-28). New York, NY: Russell Sage Foundation.

Halversen K.T. (1994). The reporting format. In H. Cooper & L.V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 425-438). New York, NY: Russell Sage Foundation.

Hedges, L.V.& Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.

*Hermesch, C.B., McIntyre, J.F., Thomas, D.D., & Berrong, J.M. (2005) Outcome assessment of the dental early acceptance program. *Journal of Dental Education*, 69(11), 1238-1241.

*Holmes, D.C., Doering, & J.V., Spector, M. (2008). Associations among predental credentials and measures of dental school achievement. *Journal of Dental Education, 72(2),* 142-152.

Hunter, J.E., & Schmidt, F.L. (2004). *Methods of meta-analysis, 2nd edition*. Thousand Oaks, CA: Sage.

Joint Commission on National Board Dental Examinations, 2009. Technical report: the National Board Dental Examinations, 2009. Chicago, IL: American Dental Association.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: Joint Committee on Testing Practices.

*Kingsley, K., Sewell, J., Ditmyer, M., O'Malley, S. & Galbraith, G.M. (2007). Creating an evidence-based admissions formula for a new dental school: University of Nevada, Las Vegas, School of Dental Medicine. *Journal of Dental Education 71(4)*, 492-500.

Kramer, G.A. (1999). Value in dental aptitude testing for minority applicants. *Journal of Dental Education 63(10):* 759-765.

Kramer, G.A., & DeMarais (1992). Setting a standard on the pilot National Board Dental Examination. *Journal of Dental Education: 56(10):* 684-688.

Kreiter, C.D., & Kreiter, Y. (2007). A validity generalization perspective on the ability of undergraduate GPA and the Medical College Admission Test to predict important outcomes, *Teaching and Learning in Medicine, 19*(2), 95–100

Kuncel, N.R, Crede, M., & Thomas, L.L. (2007). A meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (GPA) for student academic performance. *Academy of Management Learning & Education*, *6(1*), 57-68.

Kuncel, N.R., Hezlett, S.A., & Ones, D.S. (2001) A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, 127(1), 162-181.

Lewin, T. (2006, August 1), Students' Paths to Small Colleges Can Bypass SAT. *New York Times*. http://www.nytimes.com/2006/08/31/education/31sat.html?_r=1, accessed May 10, 2010.

Oh, I.S., Schmidt, F.L., Shaffer, J.A., & Le, H.,(2008). The Graduate Management Admission Test (GMAT) is even more valid than we thought: A new development in meta-analysis and its implications for the validity of the GMAT. *Management & Learning Education, 7(4),* 563-570.

Potter, R.H., & McDonald, R.E. (1985). Use and application of structural models in dental education research. *Journal of Dental Education, (49)3*, 145-153.

Ranney, R.R., Wilson, M.B., & Bennet, R.B. 2005. Evaluation of applicants to predoctoral dental education programs: review of the literature. *Journal of Dental Education, 69(10),* 1095-1106.

Reilly, R. R. & Warech, M. A. (1993). The validity and fairness of alternatives to cognitive tests. (In L.C. Wing & B.R. Gifford (Eds.), *Policy issues in employment testing* (pp. 131-224). Boston: Kluwer.)

Raudenbush, S.W., 1994. Random Effects Models in Cooper H. & Hedges, L.V. (Eds.) *The Handbook of Research Synthesis* (301-322). New York: Russell Sage Foundations

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin 1(18),* 183-192.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Rosenthal, R., & DiMatteo, M.R. (2001). Meta-Analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology,* 52:59-82. [Electronic version].Downloaded from arjournals.annualreveiws.org on 9/21/09.

* Sandow, P.L., Jones, A.C., Peek, C.W., Courts, F.J., Watson, R.E. (2002). Correlation of admission criteria with dental school performance and attrition. *Journal of Dental Education 66(3),* 385-392.

Schmidt, F.L., & Hunter, J.E. (1993). Trait knowledge, practical intelligence, general mental ability and job knowledge. *Current Directions in Psychological Science, 2*, 8-9.

Shadish, W. R. & Haddok, C.K.1994. In H. Cooper & L.V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 261-282). New York, NY: Russell Sage Foundation.

Smith, R.M., Kramer, G.A., & Kubiak, A.T., 1988. Revision of Dental Admission Test standard score scale. *Journal of Dental Education*: *52(10):* 548-553.

Sterne, J.A.C., Gavaghan, D., Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology. 53(11),* 1119-1129.

Stricker, L. J., Rock, D. A., Burton, N. W., Muraki, E., Jirele, T. J., (1994). Adjusting college grade point average criteria for variations in grading standards: A comparison of methods. *Journal of Applied Psychology*, 79(2), 178-183.

Tsai, E., 2011. Personal correspondence.

Valachovic, RW, 2008. New Dental Schools: Proceed, but appreciate that they are only one of many answers to our new challenges. *Charting Progress,* monthly newsletter *of* American Dental Association. http://www.adea.org/about_adea/Documents/2008.05%20CP.mht accessed November 15, 2008.

Weaver, R.E., Chmar, J.E., Hayden, N.K. & Valachovic, R.W. (2005), Annual ADEA survey of dental school seniors. *Journal of Dental Education. 69(5)* 599-619.

Young, J. W. (1988). Developing a universal scale for grades: investigating predictive validity in college admissions. Unpublished doctoral dissertation. Stanford University, Stanford, CA.

Zwick, R., Thayer, D.T., Wingersky, M., 1994. Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests (Research Report RR-94-32). Princeton, NJ: Educational Testing Service.