THE RESEARCH BASE SUPPORTING

THE NEW Art and Science OF TEACHING

ROBERT J. MARZANO

THE RESEARCH BASE SUPPORTING THE NEW ART AND SCIENCE OF TEACHING

by

Robert J. Marzano

April 2018

The research supporting *The New Art and Science of Teaching* (NASOT; Marzano, 2017) has a long history which includes a wide variety of types of studies across a wide variety of venues and uses. This report briefly summarizes the highlights of some of the more prominent studies regarding the model.

Early Research Basis for The New Art and Science of Teaching

The New Art and Science of Teaching (Marzano, 2017) is based on a number of previous, related works including the following:

- *A Theory-Based Meta-Analysis of Research on Instruction* (Marzano, 1998)
- *What Works in Schools* (Marzano, 2003)
- *Classroom Instruction That Works* (Marzano, Pickering, & Pollock, 2001)
- *Classroom Management That Works* (Marzano, Marzano, & Pickering, 2003)
- *Classroom Assessment and Grading That Work* (Marzano, 2006)
- *The Art and Science of Teaching* (Marzano, 2007)
- *Effective Supervision: Supporting the Art and Science of Teaching* (Marzano, Frontier, & Livingston, 2011)

Each of these works was generated from a synthesis of research and theory. For example, *The Art and Science of Teaching* includes over 25 tables reporting research on the various instructional strategies presented. These tables report the findings from meta-analytic studies and the average effect sizes computed in those studies. In all, over 5,000 studies (i.e., effect sizes) are covered in the tables within the book, which represent research over five decades. The same can be said for the other titles listed above; each contains multiple tables depicting multiple studies. Thus, one can say that the NASOT was initially based on thousands of studies that span multiple decades, and these studies were chronicled and catalogued in books that have been widely disseminated in the United States. Specifically, over 2 million copies of the books and reports cited above have been disseminated to K–12 educators across the United States and in many other countries.

Evidence-Based Studies

The resources cited above indicate that the NASOT is "research-based." In general, this means that it is based on sound extant research from the field. The studies cited in the remainder of this report were conducted on the model itself. In general, such studies are typically thought of as contributing to the "evidence base" for an instructional model.

These evidence-based studies report the relationship between the model and student learning using two metrics: (1) a standardized mean difference and (2) a correlation. A standardized mean difference is commonly referred to as an effect size (although "effect size" is technically a generic term for any measure of the relationship between two variables, thus making a correlation a type of effect size). A

correlation (*r*) has a mathematical relationship with an effect size (*d*). Specifically, the following formula can be used to transform an effect size (*d*) to a correlation (*r*):

$$r = \frac{d}{(d^2 + 4)^{\frac{1}{2}}}$$

Conversely, the following formula can be used to transform a correlation (*r*) to an effect size (*d*):

$$d = \frac{2r}{(1 - r^2)^{\frac{1}{2}}}$$

Many researchers have provided guidance as to what can be considered small, medium, and large correlations and effect sizes. Using the work of Cohen (1988) and Rosenthal (1996), Ellis (2009) provides the thresholds listed in table 1.

Table 1: Thresholds for Interpreting Effect Sizes

Metric	Small	Medium	Large	Very Large
Standardized Mean Difference (*d*)	.20	.50	.80	1.30
Correlation (*r*)	.10	.30	.50	.70

While table 1 provides useful guidance, it is important to note that some researchers have argued against classifying effect sizes and correlations by their size. For example, Glass, McGaw, and Smith (1981) explain:

> There is no wisdom whatsoever in attempting to associate regions of the effect size metric with descriptive adjectives such as “small,” “moderate,” “large,” and the like. Dissociated from a context of decision and comparative value, there is little inherent value to an effect size of 3.5 or .2. Depending on what benefits can be achieved at what cost, an effect size of 2.0 might be “poor” and one of .1 might be “good.” (p. 104)

In addressing this issue, Lipsey and colleagues (2012) explain that those interpreting effect sizes must think in terms of practical significance, which involves a comparison with typical expectations. They note:

> Practical significance is not an inherent characteristic of the numbers and statistics that result from intervention research—it is something that must be judged in some context of application. To interpret the practical significance of an intervention effect, therefore, it is necessary to invoke an appropriate frame of reference external to its statistical representation. (p. 26)

They note that appropriate frames of reference for educational interventions include expectations for normal growth, other similar interventions, and the cost and resources associated with the intervention under study.

Quasi-Experimental Classroom Studies

Beginning in 2004, data were collected at Marzano Research through classroom research projects with teachers across the country. The basic design employed by these teachers was to present the same topics to two separate, intact groups, using the same instructional activities and assessments except for a specific instructional strategy that was the focus of the study. One independent variable (treatment/control condition) was analyzed as a fixed effect. In each case, a teacher-designed pretest was used as a covariate. In effect, a fixed-effects analysis of covariance (ANCOVA) was executed for the dependent measure, which in all cases was a teacher-designed posttest. In 2009, Haystead and Marzano reported the effect sizes (*d*) in table 2.

Table 2: Effect Sizes (*d*) from Haystead and Marzano (2009)

Instructional Strategy	Effect Size (*d*)
Advance Organizers	.03
Direct Vocabulary Instruction	.44
Effort and Recognition	.31
Feedback	.10
Graphic Organizers	.29
Homework	.33
Similarities and Differences	.46
Interactive Games	.46
Nonlinguistic Representations	.38
Note Taking	.38
Practice	.32
Setting Goals/Objectives	.57
Student Discussion/Chunking	.53
Summarizing	.42
Tracking Student Progress	.87

One important perspective to keep in mind is that the effect sizes in table 2 represent how much increase an individual teacher might expect in student learning if that teacher utilized one specific instructional strategy and everything else remained constant pedagogically (for a discussion, see Technical Note). Hence, the reference point for these effect sizes is the pedagogy typically employed by a specific teacher and the students taught by that specific teacher during a specific unit of instruction. For example, consider the effect size of .46 for interactive games. This is associated with an 18 percentile point gain in student achievement. Taking the results of the studies at face value, one might conclude that if an individual teacher who does not use interactive games became proficient at doing so, the achievement of his or her students would increase by 18 percentile points on the assessments used to determine student achievement at the end of that teacher's units of instruction.

Another aspect of these findings that relates directly to the previous discussion is that in all cases teachers had minimum training in the use of instructional strategies (i.e., typically one day or less). This implies that the effects listed in table 2 require very few resources. This fact, in particular, adds to the practical significance of these findings.

A final perspective to keep in mind is that there was great deal of variation in the effect sizes within the sample. To illustrate, in their analysis of 329 effect sizes within the sample, Haystead and Marzano (2009) found that the largest positive effect size was *d* = +4.27 and the largest negative effect size was *d* = -2.40. Clearly, moderator variables were operating. Relative to this issue, there was a significant relationship between the amount of time a strategy was used and the effect size. That is, the longer a teacher used a strategy, the higher the effect size. This may well be a function of teacher pedagogical development; the more a teacher uses a specific strategy, the better he or she becomes at using the strategy.

Studies like those reported above with classroom teachers have continued since the 2009 report was published. Currently, the data base encompasses 87 schools in 26 districts, involving 509 teachers (see Marzano Research, 2018). These studies have resulted in over 1,800 effect sizes.

School-Wide Correlational Studies

Some studies have been conducted that examine the correlation between school-wide use of instructional strategies in the model and student achievement at the school level. That is, the unit of analysis in these studies was the school as opposed to individual teachers. For example, such a study was conducted in the state of Oklahoma as a part of the state department's examination of elements related to student achievement in K–12 schools (see *What Works in Oklahoma Schools: Phase I Report* and *What Works in Oklahoma School: Phase II Report* by Marzano Research, 2010 and 2011, respectively). These studies involved 59 schools, 1,117 teachers, and over 13,000 K–12 students. Collectively, the reports indicate positive relationships with various elements of the instructional model and student achievement at the school level. Using state mathematics and reading test data, 96% of the 82 correlations (i.e., 41 correlations for mathematics and 41 correlations for reading) were found to be positive, with some greater than .40. As described above, a .40 correlation (*r*) is considered halfway between medium and large and translates to an effect size (*d*) of .87 which is associated with a 31 percentile point gain in student achievement. These studies also aggregated data across nine design areas within the instructional model. All correlations were positive for this aggregated data. Seven of those correlations ranged from .33 to .40. Medium correlations were also reported for the total number of strategies used by teachers in a school, implying a school-wide effect for the use of the model. Specifically, the number of strategies teachers used in school had a .35 correlation with reading proficiency and a .26 correlation with mathematics proficiency.

Teacher-Evaluation Studies

By far, the most extensive research on the model has been conducted in the context of teacher evaluation. Specifically, the model has been employed by Learning Sciences International (LSI) as a teacher evaluation model since the implementation of Race to the Top (RTT) legislation in 2010. Currently, the evaluation model is used, to one degree or another, in 43 states. Studies that have been conducted on the model in the context of teacher evaluation fall into two broad categories: 1) studies of the overall model and 2) studies of instructional strategies within the model.

Studies of the Overall Model

In general, studies of the overall model examined the correlation between teachers' performance on the model as a whole and value-added measures (VAMs) of student knowledge gain as determined by some

year-end assessment. Such correlations are typically referred to as validity coefficients. One such study was conducted by Basileo (2016). She correlated overall average observation scores for the model with state-level VAMs. Table 3 reports results for three consecutive academic years.

Table 3: Correlations with VAMs at the Individual Teacher Level

Year	ELA	Math	Combined
2012–2013	.145 N = 10,727	.185 N = 7,192	.173 N = 13,316
2013–2014	.150 N = 10,245	.208 N = 6,750	.186 N = 12,379
2014–2015	.173 N = 9,888	.226 N = 6,624	.199 N = 12,248

N = number of teachers

The correlations reported in table 3 are based on large samples of teachers ranging from 6,624 to 13,316. Basileo made the case that these correlations are commensurate with other national models of teacher evaluation (see Kane et al., 2010).

Basileo (2016) also corrected the observed correlations for attenuation due to measurement error, which provides a better estimate of the true relationship between teacher observation scores and VAMs than do uncorrected correlations (Hunter & Schmidt, 1994). These corrected correlations are reported in Table 4.

Table 4: Correlations Corrected for Attenuation Due to Measurement Error

Year	ELA	Math	Combined
2012–2013	.195 N = 10,727	.248 N = 7,192	.232 N = 13,316
2013–2014	.201 N = 10,245	.279 N = 6,750	.250 N = 12,379
2014–2015	.232 N = 9,888	.303 N = 6,624	.267 N = 12,248

N = number of teachers

A similar study was conducted by Alexander (2016) involving school-level data as opposed to individual teacher data. These findings are reported in Table 5.

Table 5: Correlations with VAMs at the School Level

Year	Reading	Math	Combined	School Grade
2011–2012	.193 N = 1,074	.139 N = 1,074	.196 N = 1,074	.222 N = 1,074
2012–2013	.255 N = 1,093	.135 N = 1,093	.228 N = 1,093	.277 N = 1,093
2013–2014	.234 N = 955	.190 N = 955	.249 N = 955	.261 N = 955

N = number of schools

One unique aspect of Alexander's study is that it also included "school grades" in the analysis. This index is a state-level measure of the overall performance of a school relative to student outcomes.

An examination of the information in the tables 3 through 5 indicates that the majority of correlations range from small to medium. This is a strong trend for teacher observation systems used for evaluation purposes. Indeed, the Measures of Effective Teaching (MET) study funded by the Bill and Melinda Gates Foundation was an attempt to determine the correlations (validity coefficients) between teacher observation scores and VAMs for the most widely used observational systems at the beginning of the RTT implementation. The validity coefficients in the MET study ranged from .12 to .34 with an average of .22 (Bill and Melinda Gates Foundation, 2012). An important aspect of the MET study was that all observations were made by raters who were experts in the respective models they used to assign observational scores to teachers. One would naturally expect these expert raters to assign the most accurate scores possible by a human rater within the context of their respective models. Within the teacher evaluation studies on the NASOT model described above, raters were not experts in the model and, therefore, most probably had a great deal of error associated with the scores they assigned. By definition, such error will lower the observed correlations (Hunter & Schmidt, 1994). Indeed, this is the reason Basileo (2016) reported the validity coefficients in table 4, which she corrected for measurement error due to rater unfamiliarity with the technical requirements for making valid observations within the model.

The studies reported above aggregate data across districts. Some studies of the evaluation model have occurred within specific districts. Some of these studies report validity coefficients that are substantively larger than the coefficients computed across districts. For example, LSI reported that one district-specific study had a combined correlation of .332 with state VAMs. However, this correlation increased to .614 when data were limited to those teachers who actually taught subject areas directly assessed by the state test (LSI, 2013).

One district-specific study of note was conducted in Pinellas County (PC), Florida over a three year period of time starting in the 2011–2012 school year (Basileo, Toth, & Kennedy, 2015). The study examined the correlations between overall performance in the model and state VAMs at the individual teacher level. These correlations are reported in table 6.

Table 6: Correlations with State VAMs at the Individual Teacher Level

Year	2011–2012			2012–2013			2013–2014		
Subject	Read	Math	Combined	Read	Math	Combined	Read	Math	Combined
r	.168	.444	.239	.221	.460	.287	.251	.532	.347
N	61	40	75	64	41	75	64	45	75

N = number of teachers

The validity coefficients in table 6 range from small (r = .168) to large (r = .532). However, two of the nine (i.e., .287 and .251) border on medium. Thus, six of the nine validity coefficients range from bordering on medium to large. As mentioned previously, this range of validity coefficients is substantially larger than the range reported in the MET study (.12 to .33; small to medium).

In addition to examining the teacher-level correlations with state VAMs, the PC study sought to determine if training in the model produced significant increases in correlations with state level VAMs. As described by LSI (2016), during the 2012–2013 school year, Pinellas County Schools received Florida

Department of Education approval to develop teacher evaluation practices that help teachers develop their pedagogical skills. The projected outcome was to increase student achievement as teachers improved their pedagogical skills in the model. To this end, five schools were designated as treatment schools, and five matched schools were designated as control schools. Teachers' classroom practices in control schools were observed using the model, as were the classroom practices of teachers in experimental schools. In addition, teachers in the experimental schools received training in the instructional strategies within the model.

The achievement gains of students in the experimental schools were compared with those of students in the control schools. As described by LSI (2016), "Students who attended treatment schools had significantly increased growth scores (.37 to .39 standard deviations above predicted) compared to students at control schools" (p. 5). It is noteworthy that this approach was recognized by the U.S. Department of Education as an emerging approach to teacher development and evaluation (Reform Support Network, 2015).

Studies of Instructional Strategies Within the Model

One of the unique features of studies on the evaluation model is that data have been collected on teachers' use of specific strategies. Specifically, evaluation scores for teachers' use of specific instructional strategies have been correlated with state-level VAMs for specific instructional strategies (Basileo, 2016). These correlations are reported in tables 7, 8, and 9, which represent three consecutive years.

Table 7: Correlations (*r*) with 2014–2015 State-Level VAMs

2014–15 Element Correlation to VAM	ELA VAM	Math VAM	Combined VAM
Applying Consequences for Lack of Adherence to Rules and Procedures	.171	.224	.200
Helping Students Practice Skills, Strategies, and Processes	.149	.182	.173
Organizing Students to Interact with New Content	.161	.159	.172
Organizing Students to Practice and Deepen Knowledge	.151	.190	.172
Identifying Critical Content	.154	.161	.164
Probing Incorrect Answers	.146	.169	.164
Maintaining a Lively Pace	.144	.160	.160
Helping Students Record and Represent Knowledge	.139	.172	.159
Establishing Classroom Routines	.134	.191	.158
Noticing When Students are Not Engaged	.127	.190	.153
Tracking Student Progress	.133	.174	.153
Demonstrating Withitness	.132	.170	.152
Engaging Students in Cognitively Complex Tasks	.131	.127	.150
Chunking Content into Digestible Bites	.138	.131	.148
Helping Students Elaborate on New Content	.141	.133	.148
Providing Rigorous Learning Goals and Performance Scales	.128	.163	.148
Helping Students Process New Content	.131	.161	.147
Managing Response Rates	.124	.171	.146
Reviewing Content	.118	.178	.146
Helping Students Examine Similarities and Differences	.135	.138	.145
Organizing the Physical Layout of the Classroom	.130	.147	.142

Asking Questions of Low Expectancy Students	.120	.170	.141
Helping Students Revise Knowledge	.123	.139	.140
Providing Resources and Guidance for Cognitively Complex Tasks	.144	.114	.140
Organizing Students for Cognitively Complex Tasks	.091	.182	.139
Using Homework	.126	.146	.138
Previewing New Content	.126	.140	.135
Helping Students Examine Their Reasoning	.099	.155	.124
Presenting Unusual or Intriguing Information	.122	.070	.124
Helping Students Reflect on Learning	.119	.104	.122
Using Physical Movement	.094	.137	.121
Using Verbal and Nonverbal Behaviors that Indicate Affection for Students	.118	.103	.115
Providing Opportunities for Students to Talk	.111	.097	.113
Using Academic Games	.077	.152	.109
Acknowledging Adherence to Rules and Procedures	.071	.154	.102
Demonstrating Intensity and Enthusiasm	.083	.119	.097
Demonstrating Value and Respect for Low Expectancy Students	.068	.141	.095
Celebrating Success	.078	.108	.085
Using Friendly Controversy	.065^{ns}	.150	.085
Displaying Objectivity and Control	.059	.125	.075
Understanding Students Interests and Backgrounds	.047^{ns}	.101	.061

N of Teachers = 11,452

Figure 8: Correlations (*r*) with 2013–2014 State-Level VAMs

2013–14 Element Correlation to VAM	ELA VAM	Math VAM	Combined VAM
Applying Consequences for Lack of Adherence to Rules and Procedures	.149	.194	.186
Demonstrating Withitness	.156	.172	.178
Helping Students Practice Skills, Strategies, and Processes	.123	.165	.159
Establishing Classroom Routines	.115	.188	.157
Probing Incorrect Answers	.144	.135	.155
Maintaining a Lively Pace	.123	.170	.153
Reviewing Content	.117	.175	.153
Noticing When Students are Not Engaged	.123	.164	.149
Helping Students Examine Their Reasoning	.108	.153	.143
Managing Response Rates	.109	.168	.142
Providing Rigorous Learning Goals and Performance Scales	.115	.151	.140
Chunking Content into Digestible Bites	.112	.159	.139
Organizing Students to Interact with New Content	.125	.147	.139
Identifying Critical Content	.118	.150	.138
Using Homework	.122	.151	.136
Helping Students Reflect on Learning	.101	.157	.135
Organizing Students to Practice and Deepen Knowledge	.110	.142	.133
Helping Students Process New Content	.119	.123	.132
Acknowledging Adherence to Rules and Procedures	.095	.164	.131
Organizing the Physical Layout of the Classroom	.103	.152	.129

Helping Students Revise Knowledge	.117	.116	.129
Tracking Student Progress	.107	.145	.129
Demonstrating Intensity and Enthusiasm	.109	.140	.126
Previewing New Content	.109	.125	.125
Using Physical Movement	.094	.146	.122
Helping Students Elaborate on New Content	.105	.131	.119
Helping Students Record and Represent Knowledge	.094	.131	.117
Asking Questions of Low Expectancy Students	.105	.076	.103
Organizing Students for Cognitively Complex Tasks	.061	.147	.101
Celebrating Success	.078	.119	.100
Presenting Unusual or Intriguing Information	.079	.059^{ns}	.091
Engaging Students in Cognitively Complex Tasks	.074	.090	.090
Using Academic Games	.068	.116	.086
Displaying Objectivity and Control	.072	.096	.085
Helping Students Examine Similarities and Differences	.079	.077	.085
Using Verbal and Nonverbal Behaviors that Indicate Affection for Students	.074	.098	.085
Demonstrating Value and Respect for Low Expectancy Students	.073	.077	.082
Providing Resources and Guidance for Cognitively Complex Tasks	.053	.113	.082
Providing Opportunities for Students to Talk	.062	.094	.068
Understanding Students Interests and Backgrounds	.071	.058	.066
Using Friendly Controversy	.052^{ns}	.144	.064

N of Teachers = 15,452

Table 9: Correlations (*r*) with 2012–2013 State Level VAMs

2012–13 Element Correlation to VAM	ELA VAM	Math VAM	Combined VAM
Applying Consequences for Lack of Adherence to Rules and Procedures	.134	.182	.169
Noticing when Students are not Engaged	.140	.169	.162
Examining Errors in Reasoning	.115	.180	.154
Engaging Students in Cognitively Complex Tasks	.130	.119	.152
Asking Questions of Low Expectancy Students	.109	.152	.147
Providing Clear Learning Goals and Scales (Rubrics)	.123	.146	.147
Maintaining a Lively Pace	.125	.152	.146
Establishing Classroom Routines	.115	.163	.143
Reviewing Content	.124	.138	.142
Revising Knowledge	.111	.131	.136
Demonstrating “Withitness”	.109	.146	.134
Previewing New Content	.125	.115	.133
Organizing Students to Practice and Deepen Knowledge	.115	.128	.131
Practicing Skills, Strategies, and Processes	.104	.143	.131
Probing Incorrect Answers	.089	.155	.130
Reflecting on Learning	.113	.119	.130
Tracking Student Progress	.110	.124	.130
Elaborating on New Information	.105	.128	.129

Identifying Critical Information	.108	.129	.129
Managing Response Rates	.109	.132	.129
Organizing the Physical Layout of the Classroom	.105	.128	.123
Organizing Students to Interact with New Knowledge	.096	.119	.122
Processing New Information	.095	.125	.122
Chunking Content into Digestible Bites	.094	.131	.120
Celebrating Success	.100	.118	.118
Acknowledging Adherence to Rules and Procedures	.097	.102	.114
Recording and Representing Knowledge	.082	.118	.109
Using Friendly Controversy	.080	.104	.108
Using Homework	.069	.082	.101
Demonstrating Intensity and Enthusiasm	.079	.107	.097
Examining Similarities and Differences	.099	.104	.097
Organizing Students for Cognitively Complex Tasks	.062^{ns}	.108	.097
Understanding Students Interests and Background	.079	.099	.096
Providing Resources and Guidance	.067	.101	.091
Providing Opportunities for Students to Talk	.059	.096	.081
Demonstrating Value and Respect	.070	.084	.079
Displaying Objectivity and Control	.081	.039^{ns}	.077
Using Academic Games	.043^{ns}	.104	.077
Using Physical Movement	.041^{ns}	.094	.076
Using Verbal and Nonverbal Behaviors	.054	.089	.073
Presenting Unusual or Intriguing Information	.041^{ns}	.056^{ns}	.060

N of Teachers = 13,326

Again, the correlations in these tables are typically small to approaching medium in size. However, they might be considered quite large if one takes the perspective of practical significance described previously. As argued by Glass and colleagues (1988), a correlation of .10 might be considered good depending on what benefits can be achieved at what cost. For the sake of illustration, assume that the typical correlation between a single instructional strategy in the model and student learning as measured by VAMs computed from end-of-year state tests is .10 (although the majority of correlations for each year were above .10). As indicated in table 1, a correlation (*r*) of .10 translates to an effect size (*d*) of .20. An effect size of .20, as it relates to the individual strategies in tables 7, 8, and 9, implies that if teachers use a specific instructional strategy, the average achievement of their students as measured by VAMs on state-level end of year tests would be 8 percentile points higher than they would be if the teacher did not use the strategy. The practical significance of these findings is compelling in that it takes little time and virtually no financial resources for teachers to ensure that they use specific instructional strategies. Stated differently, the resource cost of using specific instructional strategies is negligible, yet this rather simple intervention can produce results (i.e., an 8 percentile point gain in average VAMs) most teachers, schools, and districts would welcome.

Distal Versus Proximal VAMs

An interesting study regarding the strength of the relationship between instructional strategies in the model and student learning was reported in "Proximal Versus Distal Validity Coefficients for Teacher Observation Instruments" (Marzano, 2014). As demonstrated above, with some exceptions, virtually all

of the correlations between teacher use of instructional strategies and student learning within teacher evaluation studies (regardless of the evaluation model that is used by a school or district) fall somewhere between small (r = .10) and medium (r = .30). As explained above, small to medium correlations are a strong trend for teacher observation systems used for evaluation purposes. Recall the average correlation of .22 between observation scores by experts and state-level VAMs reported in the MET study. While research indicates that correlations of this size between teacher observation scores and student achievement are large enough to demonstrate the validity of observations systems (Kane, Taylor, Tyler, & Wooten, 2010), they still seem rather small in the grand scheme of educational practice.

To address this issue, Marzano (2014) sought to determine if validity coefficients are higher when student learning is measured at the same time that teacher observations are made. To do so, observations were made of 79 teachers using the NASOT framework, and student learning was determined using gain in student understanding during the same lesson as determined by an increase in questions answered correctly from the beginning of the class to the end of the class. The overall correlation (i.e., proximal validity coefficient) for observation scores was .75, which, of course, would be considered very large using the guidelines in table 1. The proximal validity coefficients were compared with correlations for a sample of similar teachers and similar students in which year-end assessments were used as the criterion. These distal validity coefficients were .17, .21, and .26 for reading, writing, and mathematics VAMs respectively. These findings were taken as evidence that student effects from instructional strategies most commonly manifest during the same class period in which the strategy is used. These findings make those reported in tables 7, 8, and 9 even more noteworthy. If the effects of specific instructional strategies manifest during the same class period in which they are used, then it is somewhat striking that their use would have any effects lasting until the end of the year. One might conclude that specific strategies not only help immediate understanding on the part of students, but in some cases help anchor new content in long term memory.

Mediating Effects of the Model on Technology

One set of studies examined the mediating effects of the NASOT on the use of technology. Specifically, a two-year study was conducted to determine (in part) the relationship between selected strategies from the NASOT model and the effectiveness of interactive whiteboards in enhancing student achievement (see *Final Report: A Second Year Evaluation Study of Promethean ActivClassroom* by Haystead & Marzano, 2010). In all, 131 experimental/control studies were conducted across various grade levels. Selected strategies were correlated with the effect sizes (d) for use of the interactive whiteboards. All correlations for the strategies were positive, with some as high as .70. This implies that the effectiveness of interactive whiteboards as used in these 131 studies was enhanced by teacher use of the instructional strategies in the NASOT model.

Conclusions

The NASOT model has decades of research-based and evidence-based support for its utility in K–12 education. These studies continue to date, as the model is used in different venues and with different subject areas. However, the tens of thousands of studies conducted on the model since the turn of the century appear to indicate that the instructional strategies in the model considered in isolation, and the model considered as a whole, have a very stable, positive influence as measured by student learning in the near term and far term.

Technical Note

Interpretations of Standardized Mean Differences (*d*) and Correlations (*r*) as Effect Sizes: There are a number of resources that describe the interpretations of *d* and *r* as effect sizes in nontechnical terms (e.g., Marzano, Waters, & McNulty, 2005). Briefly, when interpreting these two metrics as effect sizes, it is important to understand the concept of a *z* score. A *z* score represents the transformation of a raw score into standard deviation units. Thus, a *z* score of 1.00 means that a given raw score is one standard deviation above the mean within the distribution of raw scores from which the observed scored is derived. A *z* score of -1.00 means that a given raw score is one standard deviation below the mean; a *z* score of 2.00 means that a given raw score is two standard deviations above the mean, and so on. This understanding provides a straightforward interpretation of the standardized mean difference (*d*) effect size.

The standardized mean difference (*d*) is most commonly used in an experimental situation, where an intervention (like the use of a specific instructional strategies or set of strategies) is examined by having an experimental group (i.e., the group in which the instructional strategy is used) and a control group (i.e., the group in which the instructional strategy is not used). Theoretically, the experimental and control groups are expected to be identical except for the fact that the experimental group uses the intervention and the control group does not. The basic formula for *d* is:

$$\text{standardized mean difference } (d) = \frac{(\text{mean of the experimental group}) - (\text{mean of the control group})}{(\text{estimate of the population standard deviation for the outcome})}$$

It is important to note that there are a number of ways to estimate the population standard deviation along with techniques for computing the effect size under different assumptions (see Cohen, 1988; Glass, 1976; Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985). Arguably the most commonly used method is that developed by Cohen and referred to as Cohen's *d*. To illustrate how an effect size is computed, assume the achievement mean of the experimental group is 90 and the achievement mean of the control group is 80. Also, assume that the population standard deviation is 10. The effect size would be:

$$d = \frac{90 - 80}{10} = 1.0$$

This effect size of 1.0 can be interpreted in the following way: The mean of the experimental group is 1.0 standard deviation higher than the mean of the control group. Assuming that the study was done rigorously, a defensible inference, then, is that the use of the instructional strategy increased students' learning by one standard deviation. Thus, the effect size *d* expresses the impact of an intervention in standard deviation units. It is this characteristic that allows one to attach a specific percentile gain (or loss) to a specific effect size expressed as *d*.

Percentile gain (or loss) is the expected gain (or loss) in percentile points of the average student in the control group as a result of the intervention used in the experimental group. To illustrate, consider the example above. Given an effect size of 1.0, one can conclude that the average score in the experimental group is 34.134 percentile points higher than the average score in the control group. This is necessarily the case since *d* is expressed in *z* score form, and distribution theory tells us that a *z* score of 1.0 is at the 84.134 percentile point of the standard normal distribution.

As described above, the effect size *r* is related to effect *d*. Using the equations above, one can transform an effect size expressed as *d* into an effect size expressed as *r*. However, the effect size *r* has a slightly more complex interpretation than does *d*. At a basic level, *r* still represents the strength of the

relationship between two variables. In all the examples used in this paper, one variable is the teachers' effectiveness at using specific instruction strategies, and the other variable is student learning. The effect size *r* can be interpreted in terms of how much one would expect student learning to increase as teachers increased their skill at using specific instructional strategies. To illustrate, consider the following formula:

$$\text{predicted } z \text{ score for student learning} = (\text{observed } z \text{ score in use of instructional strategy}) \times (r)$$

To understand the basic dynamic of this equation, it is best to start with the term "observed *z* score in use of instructional strategy." Relative to any single instructional strategy or group of strategies executed as a set, one can reasonably assume that there is a normal distribution in terms of teachers' skills. Thus, the level of expertise for any teacher on this distribution could be represented as a *z* score. If one knows the correlation (*r*) between the *z* score on the instructional expertise distribution and the *z* scores on the student learning distribution, then one can predict the typical (i.e., average) achievement of students in the teacher's class by multiplying the observed expertise *z* score by the correlation. For example, assume that the correlation between the use of a specific instructional strategy used during a unit of instruction and student achievement on a specific summative test for that unit of instruction is .50. If a specific teacher has a *z* score of 1.00 relative to her expertise in that strategy, then one would predict that the typical achievement of her students would be a *z* score of .50. Of course, a *z* score of .50 means that the average score for the students in her class would be at the 69.25 percentile in terms of achievement on the end of unit test.

Using this same logic, the correlation can be used to estimate the expected increase in student achievement, if teachers increased their pedagogical skills. This is best exemplified if one assumes that a teacher's observed expertise is represented by a *z* score of 0.00 (i.e., the teacher is at the 50th percentile in expertise) and the average score for the students in her class is represented by a *z* score of 0.00 (the class average is at the 50th percentile). If the teacher increases her instructional expertise by one standard deviation, she would now have a *z* score of 1.00. Using the formula above, one would predict that student achievement would rise to the *z* score level of .50 which would put the class at the 69.15 percentile. Stated differently, an increase in one standard deviation in teacher expertise would be associated with an 19.15 percentile increase in students' achievement relative to the end-of-unit summative test.

References

Alexander, S. (2016). *The relationship between teacher evaluation model, value-added model, and school grades*. Ft. Lauderdale, FL: Keiser University.

Basileo, L. D. (2016). *Validity coefficients results from a state-level analysis: Does the Marzano Teacher Evaluation Model compare?* West Palm Beach, FL: Learning Sciences International.

Basileo, L. D., Toth, M. D., & Kennedy, E. A. (2015). *Final report: Pinellas County Public Schools 2013–2014 multiple measures pilot results.* West Palm Beach, FL: Learning Sciences International.

Bill and Melinda Gates Foundation (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Author.

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003.

Ellis, P. D. (2009). *Thresholds for interpreting effect sizes*. Hong Kong Polytechnic University. Retrieved from http://www.polyu.edu.hk/mm/effectsizefaqs/thresholds_for_interpreting_effect_sizes2_html

Glass, G. V. (1976). Primary, secondary, and meta-analyses of research. *Educational Researcher, 5*, 3–8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Sage: Beverly Hills.

Haystead, M. W., & Marzano, R. J. (2009). *Meta-analytic synthesis of studies conducted at Marzano Research on instructional strategie*s. Englewood, CO: Marzano Research. Retrieved from https://www.marzanoresearch.com/meta-analytic-synthesis-of-studies

Haystead, M. W., & Marzano, R. J. (2010) *Final report: A second year evaluation study of Promethean ActivClassroom*. Englewood, CO: Marzano Research. Retrieved from https://www.marzanoresearch.com/2nd-year-evaluation-study-of-promethean-activclassroom

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

Hunter, J. E., & Schmidt, F. L. (1994). Correcting for sources of artificial variation across studies. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 323–336). New York: Russell Sage Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. (March, 2010). Identifying effective classroom practices using student achievement data (March 2010). NBER Working Paper No. w15803. Retrieved from https://ssrn.com/abstract=1565963

Learning Sciences International (2013). *First implementation year study for the correlation of value-added model scores, average scores on Marzano elements, and quality ratings for pedagogy*. West Palm Beach, FL: Author.

Learning Sciences International (2016). *The research base for the Marzano teacher evaluation model and correlations to state VAM*. West Palm Beach, FL: Author.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms.* Washington, DC: Institute for Education Sciences, U.S. Department of Education.

Marzano, R. J. (1998). *A theory-based meta-analysis of research on instruction*. Denver, CO: Mid-continent Regional Educational Laboratory.

Marzano, R. J. (2003). *What works in schools*. Alexandria, VA: ASCD.

Marzano, R. J. (2006). *Classroom assessment and grading that work*. Alexandria, VA: ASCD.

Marzano, R. J. (2007). *The art and science of teaching*. Alexandria, VA: ASCD.

Marzano, R. J. (2014). Proximal versus distal validity coefficients for teacher observational instruments. *The Teacher Educator, 49*(2), 89–96.

Marzano, R. J. (2017). *The new art and science of teaching*. Bloomington, IN: Solution Tree Press.

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching.* Alexandria VA: ASCD.

Marzano, R. J., Marzano, J. S., & Pickering, D. J. (2003). *Classroom management that works*. Alexandria, VA: ASCD.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works*. Alexandria, VA: ASCD.

Marzano, R. J., Waters, T., & McNulty, B. A. (2005). *School leadership that works*. Alexandria, VA: ASCD.

Marzano Research. (2010). *What works in Oklahoma schools: Phase I report*. Englewood, CO: Marzano Research.

Marzano Research. (2011). *What works in Oklahoma schools: Phase II report*. Englewood, CO: Marzano Research.

Marzano Research. (2018). *Meta-analysis database of instructional strategies*. Centennial, CO: Author.

Reform Support Network. (2015). Emerging approaches to measuring student growth. Retrieved from https://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/emergapprotomeasurstudgrowth.pdf

Rosenthal, J. A. (1996). Qualitative descriptors of strength of association and effect size. *Journal of Social Service Research, 21*(4), 37–59.

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research: Methods and data analysis.* New York: McGraw-Hill.