Running head: COGNITIVE TRANSFER FROM ARTS EDUCATION

Cognitive Transfer from Arts Education to Non-arts Outcomes:

Research Evidence and Policy Implications

Lois Hetland

Project Zero, Harvard Graduate School of Education

Ellen Winner

Department of Psychology, Boston College and Project Zero, Harvard Graduate School of Education

In E. Eisner and M. Day (Eds.), Handbook on Research and Policy in Art Education. National Art Education Association, 2002.

Abstract

This chapter reports methods, findings, and implications for research and policy from 10 meta-analytic reviews of the effects on non-arts cognition from instruction in various art forms. Three analyses demonstrate generalizable, causal relationships: classroom drama and verbal achievement, music listening and spatial reasoning, and music learning and spatial reasoning. Five do not allow causal conclusions: multi-arts and academic achievement, arts rich instruction and creativity, visual arts and reading, dance and reading, music and reading. Findings for two analyses are equivocal: dance and spatial reasoning, music and mathematics. The authors urge arts education researchers to keep research syntheses in mind when conducting studies and advise policy-makers to support arts programs that demonstrate learning in the arts.

Cognitive Transfer from Arts Education to Non-arts Outcomes: Research Evidence and Policy Implications

Both researchers and policy-makers in arts education seek understanding, yet they differ in what satisfies that quest. Researchers pursue enduring puzzles--puzzles that expand to require new insights from the next study, or even from studies conducted by later generations of researchers. Following the trail of evidence wherever it leads, research advances incrementally, often over the course of decades. Scholars develop patience with this glacial pace, with one study leading to another as they probe another nuance, a further connection. Policy, on the other hand, cannot afford patience. Policy-makers need to act, and they need to act now. Children grow up, money is allocated, and programs are implemented, always today. The urgency of policy contrasts starkly with the slowly accruing clarity developed through the enterprise of research.

Despite these fundamentally different drives, both researchers and policy-makers can benefit from the research produced by REAP (Reviewing Education and the Arts Project, Winner & Hetland, 2000): quantitative syntheses of individual studies that test the effects of arts education on learning in non-arts domains. While the latest study may seem like the answer for awhile, its allure and veracity fades as the next new study rises on the horizon. No matter how well designed, single studies offer only partial answers, particular to the settings, persons, and procedures of their design. In contrast, the lens provided by looking at the evidence amassed from a body of studies on a given question can clarify apparently contradictory evidence and allow clear patterns to emerge that can guide further research as well as practical applications in policy. By combining and comparing findings from all available studies that address similar questions, syntheses help researchers understand variation, develop better methods, and identify

new questions. The same summaries offer policy-makers what they need most--the best evidence available at a given time upon which to base decisions (Light & Pillemer, 1984).

In today's educational climate, academic skills seem to be valued exclusively, and all too often the arts are seen as expendable frills. In such an environment, arts advocates need to convince decision makers of the rightful place of the arts in the schools. But as they look to research to build their case, they find scattered evidence, and much of it is not about the inherent value of the arts for children, but rather about the instrumental value of the arts—their effects on basic academic skills whose importance is undisputed.

Arts educators and advocates have argued with increasing fervor over the past two decades that the arts are a means to improved basic academic skills. For example, according to a 1995 report by the President's Committee on the Arts and Humanities, "teaching the arts has a significant effect on overall success in school " (Murfee, 1995, p. 3). The report justifies this claim by noting that both verbal and quantitative SAT scores are higher for high school students who take arts courses than for those who take none. And former Secretary of Education, Richard Riley, stated that "The arts teach young people how to learn by giving them the first step: the desire to learn" (Fiske, 1999, p. vi).

But what is the research base on which such claims are made? These claims were not based on summaries of the research evidence for transfer of learning from the arts to academic subjects, because no summary had been conducted to assess the strength of the case or to understand the mechanisms of transfer, who benefited, and under what conditions. We therefore set out to identify and synthesize all studies on this question since 1950. In this chapter we describe our methodology and present the results of ten quantitative syntheses. Three of our syntheses revealed a clear case for transfer. In two syntheses, the claims are equivocal. In five,

we had to conclude that there is (at least as yet) no compelling evidence that study in an art form leads to improved academic functioning. At the conclusion to this chapter we answer critics who have misinterpreted us as arguing that the arts therefore do not help children. We argue instead that the arts have great value in a child's education but that this value is due first and foremost to the importance of learning in the arts. While arts study may in some cases instill skills that strengthen learning in other disciplines, arts programs should never be justified primarily on what the arts can do for other subjects.

Synthesizing Evidence through Meta-analysis

The REAP reviews were conducted using a technique called meta-analysis, a quantitative synthesis in which the unit of analysis is the study rather than the person. Although meta-analysis is only now becoming commonplace in educational research, it is the standard for synthesis in public health, medicine, epidemiology, psychology, and agriculture (where it was first used). Meta-analyses bring coherence to research domains, and consequently they are among the most cited forms of research (Cooper & Hedges, 1994). Policy-makers, especially, should benefit from understanding how meta-analyses are conducted so that they can weigh the quality of two research syntheses that review the same information.

A meta-analysis can do several things that a traditional narrative review of the literature cannot. First, a meta-analysis can tell us the average strength of the relationship between arts and academic outcomes derived from many studies. That information is preferable to generalization from a single study (which by its nature is particular to the settings, procedures, and subjects it assesses). It is also preferable to merely counting the number of studies achieving significance at a level of, say $p < .05$ (which is misleading because the same effect could be statistically significant or not depending only on how many subjects were included in an analysis; studies

with larger sample sizes generate more effects judged as "significant"). Second, a meta-analysis tells us how reliable (i.e., how likely to be reproduced in next studies) this average effect size is. Meta-analyses also allow us to test hypotheses: by coding studies for important variables (e.g., length of arts study, or whether arts were taught separately or integrated into the curriculum) and then comparing average effects for the different groups, we can see the influence of particular variables. Thus, a meta-analysis is far more than a summary--it makes it possible to generate and explore new hypotheses that cannot be explored in single studies (such as whether particular research designs demonstrate larger effects than others, or whether published studies show larger effects than unpublished ones).

While the fundamental goal of meta-analytic procedures is to cumulate evidence that aids understanding of past research and guides future inquiry, meta-analytic methods are similar to other types of quantitative research. Methodological standards for meta-analytic reviews are summarized in *The Handbook for Research Synthesis* (Cooper & Hedges, 1994). Along with texts by meta-analytic specialists such as Rosenthal (1991; 1995), Light (Light & Pillemer, 1984), and Mosteller (Mosteller & Colditz, 1996), the *Handbook* served as the methodological foundation for REAP.

Meta-analysts first define a research question and a sampling frame for a population of studies (i.e., they set principles about the kind of studies to be included in the analysis). Then the search for studies (the "subjects" of a meta-analysis) begins with the aim of finding an unbiased sample of studies that fairly represents all the studies conducted on the relevant research question. Next, studies are coded objectively for potential moderator variables that may influence the effect size resulting from different "treatments" (in this case, different forms of arts

education). These might include, for example, length of time of arts study, quality of arts class, level of teacher expertise, genre of art studied, study design, or outcome measure employed.

When studies do not routinely report effect sizes (as is all too often the case), analysts must first compute effect sizes from reported data such as means, sample sizes, and significance levels. Each study contributes one average effect and one significance level to the group average, computed through standard statistical procedures. In further analysis, groups of studies can be compared to test whether moderator variables influence the sizes of average effects.

In the work described in this chapter, we first searched for all studies, published and unpublished, carried out since 1950, which examined the relationship between arts study and academic achievement. In our search, we used seven computer data bases, reviewed reference lists of acquired articles, contacted over 200 scholars in the field, and hand-searched 41 journals from the previous 50 years. Such an exhaustive, systematic search with redundant channels is the best way to identify the "fugitive" literature, which helps to minimize a common threat to validity of meta-analytic findings from sampling bias. This kind of search also reduces the likelihood of bias caused by combining only published studies. Published studies are generally "significant," since researchers often do not submit studies for publication when they have not demonstrated significance at $p < .05$ (i.e., the level of likelihood at which, were the study conducted 100 times, the reported effect would be achieved in 95 of the trials). Including a disproportionately high number of published articles in meta-analyses may artificially raise average effects and/or combined significance levels, leading to inflated, inaccurate results (Rosenthal, 1994).

From those articles identified in the REAP search we included only studies that actually quantified (a) some kind of non-arts, cognitive outcome, (b) results from subjects who received

some type of arts instruction or exposure, (c) results that compared subjects who received arts to subjects in a control group who received either no treatment or, better, an alternate treatment. Because of the volume of studies, the REAP analyses could not examine the evidence for the claimed social outcomes of the arts such as improved academic motivation or increased school attendance--studies addressing these questions remain to be synthesized. Nor did we include studies in which teachers expressed their belief that students' cognitive skills were boosted by the arts (since this was not a direct measure of cognitive improvement), nor studies that showed improvement from arts study without comparison to a control group who received no arts study over the same period of time. We found almost 200 studies that met our criteria for inclusion. We then defined criteria for inclusion for the ten separate meta-analyses and sorted the studies accordingly.

Since meta-analyses combine only quantitative findings, it is worth a brief digression at this point in the description of our methods to discuss whether our results represent an unbiased sample of the research conducted on questions of arts transfer. To be included in a meta-analysis, qualitative results must include numerical values (e.g., as they did in Heath's [1998] qualitative study, which was included in the Winner & Cooper [2000] analysis summarized below). If they do not, they cannot contribute directly to a meta-analytic average effect size. However, that fact neither reduces the contribution of qualitative studies to meta-analyses nor the value of meta-analyses as unbiased summaries. Qualitative data contribute to meta-analytic reviews in their discussion sections, not in the quantified results. That is because, in the ongoing investigation of psychological and behavioral phenomena, qualitative research contributes different kinds of evidence than its quantitative cousins. Qualitative research does not attempt to calculate the size of a relationship, but rather seeks to inform by triangulating examples of phenomena observed in

lived contexts, by "thick" description, and by categorical analysis of the nature and dimensions of relationships. Ideally, qualitative and quantitative approaches to a given question build upon one another, focusing on the problem first in one way, then another, as befits particular puzzles. In review and summary of a field, qualitative data contribute observations that inform interpretations of numerical results and contribute hypotheses that can be tested quantitatively--either within the qualitative study itself or in related quantitative study later on.[1] Thus, qualitative studies located by REAP's exhaustive search, as well as those studies rejected because they did not meet study inclusion criteria, should not be expected to contribute to the quantitative cumulation of effects. But they do help to explain why the numerical findings turned out as they did. They suggest avenues and techniques for future investigation.

After locating studies through our search procedures, REAP researchers classified them by both art form and non-arts outcome (e.g., visual arts and reading, music and mathematics). For each set of comparable studies, we developed codes for potential moderator variables and conducted descriptive, inferential, and interpretive analyses. Descriptive analysis yields a description of the characteristics of the studies in the sample, with the most important of these being an average effect size from combining the studies. Along with the average effect, we reported the range of effects (largest and smallest), the quartiles (25^{th}, 50^{th}--also called the median, and 75^{th} percentiles), and the percentage of effects greater than zero.

Inferential analysis shows how likely results of the synthesis are to generalize to next studies. When reading the summaries, note the 95% confidence intervals around the average effect size. When these intervals span zero, confidence that a positive effect actually exists is typically reduced. Another inferential statistic is also very useful to policy-makers—the "*t* test of the mean *Zr*." This test makes it possible to determine how generalizable the results are to future

studies on the same research questions-- higher confidence is indexed by lower *p* levels associated with tests of the mean *Zr*.

Interpretive analyses help us to assess to whom the results apply and under what conditions. Our interpretive analyses consisted of contrast tests to assess which features of programs and research design influence the size of the reported effects.

The heart of a meta-analysis rests on the calculation of an "effect size" for each study, because these individual effect sizes allow us to calculate the size of effects when combined and compared across studies. Effect sizes show the strength of the relationship between two variables, in this case, between some type of arts study and some cognitive or academic outcome. Two of the most common statistics used to show effect size are *d* and *r*. Because simple algebra can readily transform one index to another (e.g., an *r* at low values such as those found in the REAP analyses can often be doubled to get a rough estimate of an equivalent *d*), the choice of effect size statistic is mainly based on ease of interpretation for a particular audience and on the ability to compute the chosen statistic for the body of primary studies being reviewed.

In the analyses reported here, we used *r*, as recommended by Rosenthal (1991, 1994; Rosenthal & Rosnow, 1991). This statistic can be used even when a study has more than two conditions, as did many of the studies analyzed by REAP. In contrast, *d*, which is the standardized difference between two means, is meaningless for experiments employing three or more conditions, because differences cannot be computed for more than two groups. The effect size *r* is also readily interpreted by readers unfamiliar with statistics, since it can be translated into percentages of subjects helped or not helped by a treatment. Rosenthal (1991) demonstrates this translation using a binomial effect size display (BESD) (Rosenthal & Rubin, 1982; Rosenthal, Rosnow, & Rubin, 2000).[2]

Effect size *r*s range from –1.0 (a perfectly correlated negative effect) to +1.0 (a perfectly correlated positive effect). In general, meta-analysts interpret *r*s of .10 as small, of around .24 as medium, and above .37 as large effects (Cohen, 1988). However, these are rules of thumb defined statistically, not by influence on the field, so these quantities should not rigidly direct interpretations about the importance of any given effect. Small effects sometimes matter a lot (e.g., if they index a small number of students who stay in school as a result of a treatment), and sometimes very little (e.g., if they index a rise of a few points on a standardized test). A classical example of an important small effect comes from Smith, Glass, & Miller (1980), in which psychotherapy yielded $r = .32$, and consequential effects from biomedical research are often much smaller--even as low as $r = .034$ (Steering Committee of the Physicians Health Study Research Group, 1988).

As stated by Rosenthal and Rosnow (1991), the relationship between level of statistical significance and effect size can be understood as follows:

Significance Test = Effect Size x Study Size

Simply put, the larger a study's sample size, the more significant the results will be.[3] The same small, moderate, or large effect size could be significant or not *depending only on the size of the sample*, and this is frequently forgotten in the interpretation of research results. Thus, complete reporting of the results of any study requires reporting both effect size and level of significance. All too often, studies report only half the necessary information--how likely an effect would be to occur again (the *p* value). But without knowing how large the effect is, we cannot judge the importance of its likelihood of re-occurrence. Reporting effect sizes is becoming standard practice in other fields (see the APA task force on statistical significance, Wilkerson, 1999), and arts education research publishers should require it.

In summary, we sought to employ standardized methods to ensure reliability (i.e., would replication of our analysis result in similar answers?) and validity (i.e., did we analyze what we intended to?). We do not claim that REAP is the final word on the effects of arts on cognitive transfer to non-arts areas. As more research accrues, the evidence may well require different conclusions. But we venture to assert that our findings represent the most trustworthy knowledge currently available about the question of cognitive transfer.

What the Meta-analyses Revealed

Support for Three Instrumental Claims

As noted above, our findings revealed three areas in which a causal relationship between arts and some non-arts cognitive outcome was demonstrated: classroom drama and verbal achievement; music listening and spatial reasoning; and music instruction and spatial reasoning.

Classroom Drama and Verbal Skills. Perhaps the most well-researched arts to academics transfer literature focuses on the effects of "classroom drama." Classroom drama refers to using acting techniques within the regular classroom curriculum. This stands in contrasts to theater, or the production of plays. In an earlier synthesis of this area, Kardash and Wright (1986) meta-analyzed 16 studies of classroom drama and found positive relationships between drama and reading, oral language development, self-esteem, moral reasoning and various drama skills (with an average effect size of r=.32, equivalent to d=.67). A second meta-analysis was conducted by Conard (1992) on the effect of classroom drama on verbal achievement, self-concept, and creativity. This analysis combined 20 studies, six of which were included in Kardash and Wright's analysis. Again a positive effect was found, with an average effect size of r=.23 (equivalent to d=.48). Neither of the two previous meta-analyses teased apart specific components of classroom drama that might influence academic achievement. Nor did these

previous studies separate the different kinds of outcomes that were affected and so were not able to determine which area or areas of academic achievement were more strongly related to classroom drama.

Podlozny (2000) found almost 200 experimental studies probing the effect of classroom drama on academic achievement. Over 40% of these studies tested verbal achievement outcomes, and it is this body of literature that Podlozny examined meta-analytically as part of REAP. Eighty studies were included in her meta-analysis. The studies tested and compared the effect of classroom drama on seven distinct verbal outcomes. To test which instructional qualities might influence the size of effect, Podlozny identified and assessed three components of classroom drama—*enactment*, *plot* (level of structure), and *leader* (teacher's level of involvement).

By definition, dramatic instruction entails enactment of some pretend, imaginary situation. But the form of enactment can range widely. Stories can be enacted by creating dialogue, for example, while sitting in a circle on the floor (called *verbal enactment* by Podlozny), or through pantomime (*physical enactment*). Further, enactment may be performed by the child *(self enactment*) or through puppets, toys, or other objects (*distanced from self*). Enactment often involves combinations of these four features (verbal/action/self/ distanced from self). Because 75 of the 80 studies engaged children in verbal, physical, self-drama, Podlozny was unable to test the effect of type of enactment, but such an examination merits future attention.

Plot may be *structured*, as when children are given a story or script, relatively *unstructured*, as when children are simply given themes to act out, or combine structured and unstructured plots.

The *leader* (i.e., teacher) can take the part of a character (*in-role*), work as a coach outside of the dramatic frame (*facilitator*), or work at a distance from the action (*removed*), answering questions, but not serving as a driving force for the activity.

Podlozny classified studies in terms of whether they directly tested material students had actually enacted in their drama sessions (*direct*) or whether tests were of entirely new material (*transfer*). This distinction was made to determine whether enacting a story simply helped children better read, understand, and recall a particular story that they had acted out, or whether the experience of acting out a story helped children's verbal skills more generally.

Podlozny examined seven verbal outcomes:

In 17 studies with oral recall outcomes, the drama group heard and enacted the stories and the control group heard but did not act out the stories. Students were then tested on oral recall.

In 14 studies with written recall outcomes, the drama group read and then enacted the stories while the control group read, then discussed, and were drilled on vocabulary from the stories. Children were tested only on stories that had been taught.

In 20 studies with reading achievement outcomes, the drama group typically read a story or play and enacted it while the control group simply continued with their regular reading classes. Both groups were then given a standardized reading comprehension test. Thus in this body of studies children were always tested on new material. Hence, any effect demonstrates transfer of reading comprehension skills to new material.

In 18 studies with reading readiness outcomes, the drama group heard a story and acted it out, while the control group either heard the same story and discussed but did not enact it, re-enacted themes from field trips or other experiences (and hence did not hear the story), or

engaged in cut and paste and categorizing activities (here they neither heard the story nor engaged in any enactment). This body of studies again only tested children on new material.

In 20 studies with oral language development outcomes, students in the drama group typically engaged in creative dramatics (storytelling, role-playing, puppetry) as well as discussion while the control group watched filmstrips and engaged in arts other than drama. Later the oral language of all children was assessed, sometimes when talking about new material, other times when talking about the stories that they had enacted.

In 10 vocabulary studies, children in the drama group engaged in creative drama activities, including role play, pantomime, movement, and improvised dialogue, while the control group had no special treatment. Later all children were given a vocabulary test, sometimes with words from the stories that had been taught and other times with new words.

In eight studies with writing skills outcomes, writing samples were assessed for skills such as audience awareness, story structure (beginning, middle, and end), organization, and elaboration. Typically children in the drama group first participated in a discussion about writing, and then engaged in improvisation, pantomime, and movement, developed story ideas, improvised story scenes, and drafted stories. The control group also participated in a discussion about writing, but then they simply continued with their regular language arts program before drafting their stories. Stories were analyzed according to a narrative writing scale. In some of the studies, children wrote stories related to themes they had enacted. In others, they wrote stories on new material.

Classroom drama was found to have a strong positive effect on six of the seven verbal outcomes examined. The largest effect size was for written story recall, where an average effect size of r=.50 was found (equivalent to d = 1.15; 95% confidence interval was r = .37 to r = .73; t

test of the mean $Zr = 3.91$, $p < .0025$). This is an extremely large effect. Studies assessing the effect of drama on oral language were also moderate to large (r=.30, equivalent d =.63), followed by story understanding as measured orally, reading readiness, writing, and reading achievement (r= ..27, .25, 24, .20, respectively, equivalent to d =.56, .52, .49, .41). All of these effects were robust: t tests of their mean Zrs indicate that the results generalize to future studies, and none of the confidence intervals spanned zero. Vocabulary was also enhanced (r=.06, equivalent d =.12), but unlike the other six effect sizes, this one was not statistically significant (the 95% confidence interval r = -.07 to r = .19 spanned zero, and the t test of the mean $Zr = 1.01$, $p < .24$).

The type of plot used in the drama instruction influenced the effectiveness of the instruction. Working with structured plots resulted in larger average effects when story comprehension and structure were the outcomes. When oral language development was the outcome, working with unstructured plots in improvised role-play (or combinations of structured and unstructured plots) resulted in larger average effects. Podlozny explained that because oral language development is not directly related to story structure and comprehension, this variation can be readily explained. Emphasis on extemporaneous and improvised speech is more facilitative of oral language development than working with scripted plots, which emphasize acting within the confines of the particular story or script. Interestingly, structured and unstructured plots were equally effective for oral measures of story understanding and vocabulary.

Level of teacher involvement could only be investigated as a factor in studies assessing story understanding, as none of the other studies described the role of the teacher. This factor proved to be related to the effectiveness of drama instruction for studies measuring story understanding. Following Dansky's (1980) "multi-stage" model of effects, Podlozny had

hypothesized that leader “in-role” might increase the occurrence and/or quality of dramatic play, which, in turn, might increase academic achievement.

For the question of whether drama helps students with new texts, Podlozny's seven analyses demonstrate higher effect sizes for material studied directly. However, to a lesser but still impressive degree, the analyses also show that drama helps learners understand new texts. As Podlozny says, "What is remarkable is not that drama's strongest effects are direct ones, but rather that drama does have the power to foster skills that then transfer to new material" (Podlozny, 2000, p. 266).

With respect to the age at which drama is most likely to result in enhanced verbal skills, the evidence was inconsistent. While the meta-analyses by Kardash and Wright and by Conard both found that classroom drama was more effective for younger children, five of the seven meta-analyses performed by Podlozny (2000) showed no relationship between age and effect size. Of the remaining two, one showed that the effect was stronger for younger children (writing achievement), while the other found that the effect was stronger for older children (oral language).

Also contrary to both of the previous meta-analyses, five of Podlozny's seven meta-analyses found that drama was equally effective for average, low SES, and learning disabled students. The remaining two analyses (those assessing written story understanding and reading achievement) found that drama was actually more effective in promoting verbal skills when the children involved were from low SES populations. This finding is consistent with Smilansky (1968) who reports that exposure to drama increases the achievement levels of poor students. One explanation is that children from disadvantaged backgrounds may not have engaged in as much creative, dramatic play, nor experienced success through participating in engaging

instruction. Classroom drama instruction may provide a "boost" to these students, helping them to acquire a deeper level of story understanding.

The most important finding of these meta-analyses on classroom drama is the demonstration that drama not only helps children to master the texts they enact, but also often helps them to master new material not enacted. The transfer of skills from one domain to another is generally not thought to be automatic: it needs to be taught (Salomon & Perkins, 1989). In the field of classroom drama, however, transfer appears to be naturally designed into the curriculum, even if teachers are not labeling it as such. If teachers of classroom drama did more to teach explicitly for transfer, these effects might be even stronger.

Music Listening and Spatial Reasoning. Since the prestigious journal, *Nature*, published the report by Rauscher, Shaw, and Ky (1993) that described a temporary increase in spatial reasoning by college students after listening briefly to certain kinds of music, the public has been beset by innuendoes from that now famous letter. Did the results support then Georgia Governor Zell Miller's decision to give all Georgia newborns a CD of classical music? Or Florida's mandate that day care centers should play thirty minutes of classical music daily? Or that Mozart CDs marketed to children could improve mathematics scores later in life if listened to carefully? They did not. The leap from this single laboratory experiment to policies for children's learning and to marketing strategies for prenatal courses and CDs was unwarranted and never supported by the researchers. But does that make the study or the numerous replication attempts of no interest? Indeed not--they are of great scientific interest, because they suggest that the mind may work in ways we had not previously thought.

Hetland (2000a) identified 36 relevant experiments (2,469 subjects) that could be synthesized through meta-analysis. The analysis included experiments conducted in laboratory

settings with adults (i.e., college students) who listened briefly to a musical stimulus that was predicted to enhance spatial reasoning compared to at least one control condition predicted not to enhance spatial reasoning. Replication studies also employed some measure of spatial reasoning and made enough data available to compute an effect size (i.e., the degree of relationship between the musical condition and the score on the outcome measure).

None of the replications reproduced the original experiment exactly. Many used the same music as the 1993 experiment (i.e., the *Allegro con spirito* from Mozart's Sonata for Two Pianos in D major, K. 448), but some researchers predicted that enhanced spatial reasoning would result from listening to other movements or pieces by Mozart, other classical music (e.g., Schubert, Mendelssohn), a piece by the contemporary composer, Yanni, and musical stimuli comprised only of pure rhythm or pure melody.

Replications also varied by the measures used to index spatial reasoning. "Spatial reasoning" is a term that encompasses a range of intellectual processes, much as the term "heart attack" refers to a variety of medical traumas. The Paper Folding and Cutting subtest of the Stanford-Binet: Fourth Edition, a task used in the original experiment and in many of the replications, is a good example of the type of task Rauscher and Shaw call "spatial-temporal" and which they predicted would be enhanced by listening to certain types of music. A sample item is shown in Figure 1. This task requires subjects to imagine folding and cutting paper in ways similar to actually folding and cutting paper snowflakes.

Researchers attempting to replicate the original experiment used a variety of other tests as well, some of which do not meet Rauscher and Shaw's (1998) criteria for spatial-temporal tasks. For example, Matrices tasks do not qualify as spatial-temporal. Figure 2 shows a sample item. In these tasks, one figure is missing from a gridded pattern of figures, usually 3 by 3, ordered

vertically and horizontally according to rules of logic (add the figures, subtract the figures, enlarge the figures in specific ways). Such tasks do not require flipping and turning objects mentally, nor doing so in sequential steps. The Pattern Analysis subtest of the Stanford-Binet also does not qualify as spatial-temporal, because, while it requires mentally flipping and turning objects, it provides subjects a model to match and compare while solving the task. Figures 3a and b presents a sample item of the Pattern Analysis subtest.

Another important variation in replications is the different types of control conditions employed (note that experiments often used more than one). These included silence (used in about three-quarters of the experiments), audio tapes of verbal instructions designed to lower blood pressure (used in about half of the experiments), natural and man-made sounds (5 of 36 experiments), texts read aloud (3 of 36 experiments), and music that researchers thought was not complex enough or sufficiently like Mozart to enhance spatial-temporal skills (used in about one-fourth of the experiments). For example, five of 36 experiments used a piece by Philip Glass called "Music with Changing Parts," which is almost hypnotically repetitious, and others used various "relaxing" music (one was described as "angelic female voices"). Still others used disco and rock music, presumably thought to be distracting.

Six preliminary analyses determined whether experiments with such diverse controls could be combined responsibly into a single analysis. The first two preliminary analyses replicated and compared analyses to a previous meta-analysis (Chabris, 1999). The other four preliminary analyses included studies that employed more than one control to directly compare scores on spatial-temporal tasks following different control conditions (Silence versus Relaxation tapes, Silence versus Noise, Silence versus Nonenhancing music, Relaxation versus Nonenhancing music).

The Music versus Silence analysis yielded a moderately sized average effect of $r = .24$ (equivalent $d = .48$), compared to the small average effect Chabris found, which was equivalent to $r = .07$ ($d = .14$). The Music versus Relaxation analysis yielded a moderate to large average effect of $r = .33$ (equivalent $d = .70$), compared to Chabris's similarly-sized average effect equivalent to $r = .29$ ($d = .57$). Because Hetland's sample is more representative of all the studies conducted on this research question (due to the exhaustive nature of the search and including both published and unpublished studies), these results are more likely to represent the true effect size for the theoretical "universe" of studies on this research question.

Note that the relative size of the effects for the first two preliminary analyses is similar to Chabris's analysis (i.e., Music versus Silence has a smaller effect than Music versus Relaxation). At face value, this finding lends support to the arousal theory, according to which music enhances spatial performance because it arouses. Unless over-stimulated, an aroused person performs better on tests; relaxation is likely to produce lower arousal than merely sitting in silence. However, the third preliminary analysis suggests that arousal does not account for the difference in effect sizes, because when scores following silence and scores following relaxation were compared directly, they were essentially the same ($r = -.02$, with the negative sign indicating that scores following relaxation were trivially higher on average, not lower). The remaining preliminary analyses suggest that differences in scores following various control conditions when directly compared were not consequential or systematic (Silence versus Noise, $r = .02$; Silence versus Nonenhancing music $r = -.05$; Relaxation versus Nonenhancing music, $r = -.02$). As a result of these analyses, the various control conditions used in the experiments appeared to produce essentially similar results and, thus, could be combined legitimately into a single analysis. Including all the identified experiments lends Hetland's analysis considerable

statistical power and summarizes all the laboratory data with adults identified as relevant to the question about music's temporary enhancing effect on spatial task performance.

The first main analysis (36 experiments, 2,469 subjects) compared tasks that qualified as spatial-temporal (31/36) to other types of spatial measures (5/36). Contrast analysis showed that the moderately-sized and highly generalizable mean effect (r = .22, d = .46, 95% confidence interval r = .14 to r = .31; t test of the mean Zr = 5.34, $p < .0001$) results from higher effect sizes in experiments using spatial-temporal measures. The average of the experiments employing spatial-temporal measures alone is r = .20. Experiments employing only nonspatial-temporal measures yielded an average effect of r = .04, and experiments that used a combination of spatial-temporal and nonspatial-temporal measures showed an intermediate effect size (r = .15) Thus, this analysis supports the conclusion that music's influence is specific to spatial-temporal, rather than to all types of spatial measures. Such specificity is evidence against the general arousal hypothesis.

The second main analysis included only those 31 experiments (2,089 subjects) that employed spatial-temporal measures. Again, the analysis showed a moderately-sized relationship between listening briefly to music and enhanced performance on spatial-temporal measures (r = .25, d = .50), which is highly generalizable (95% confidence interval: r = .14 to r = .35; t test of the mean Zr = 4.84, $p < .0001$). However, two problems limit the strength of the conclusions that can be drawn from the analysis.

First, the effect sizes of the individual studies varied too much to be considered as sampled from a single population of studies (Range: r = -.20 to r = .67, SD = .25, heterogeneity test, χ^2 (30) = 101.90, $p < .0001$), and only some of the variation could be accounted for by moderator variables. Of the seven potential moderator variables identified, four did not influence

the size of effect significantly (type of enhancing music used, subject gender, carry-over from previous spatial activation, and publication status). The remaining three did explain some of the variation. Experiments that employed a Relaxation tape control did have larger than average effect sizes (r = .34). However, since the third preliminary analysis showed no difference in scores following Silence and Relaxation when compared directly, it is likely that unidentified procedures of the laboratories that used relaxation as a control account for the systematic differences in effect sizes, rather than the control condition itself. This conclusion is affirmed by the results of a sensitivity analysis that temporarily removed studies from labs that contributed five or more experiments with relaxation controls. Both the Rauscher studies (average r = .40; Rauscher, Bowers, & Kohlbeck, 1999; Rauscher & Hayes 1999; Rauscher & Ribar, 1999; Rauscher, Shaw, & Ky, 1993; 1995) and the Rideout studies (r = .42; Rideout, Dougherty, & Wernert, 1998 [experiments 1 and 2]; Rideout, Fairchild, & Urban, 1998; Rideout & Laubach, 1996; Rideout & Taylor,1997) had higher than average effects.

Such an observation leads to speculation about the procedures used by various labs. An analysis of study quality showed that experiments with stronger designs (that is, designs that were less vulnerable to threats of internal validity) had higher average effects, and both the Rauscher and Rideout experiments ranked average or above on these criteria. Thus, the variation in effects is unexplained by study quality and cannot be attributed to errors by the researchers. The most likely explanation for the effect is that these two laboratories emphasized to subjects the importance of attending closely to the music. It is possible that doing so allowed the music to have an effect, while other experimental procedures allowed subjects' attention to wander. Colwell (2001) references a literature in music education that supports a conclusion that focused

attention produces a different cognitive response than does casual listening. Such an explanation should be addressed in the design of future studies.

The second limitation is that a mechanism could not be unequivocally identified as causing the effect. Experiments did not provide enough data to explore plausible alternate hypotheses to the "trion" priming model proposed by Leng and Shaw (1991), alternatives such as arousal, preference, or mood as causal mechanisms, or the theory that the element of rhythm links musical and spatial processes (Parsons et al., 1999), or the possibility that musical sophistication and training result in listening analytically and increasing the effect.

In summary, the synthesis of the "Mozart effect" studies is of scientific interest, because the highly significant, moderately-sized effect indicates that a relationship does exist between musical and spatial reasoning, as far as can be assessed from the studies conducted to date. It appears that spatial and musical processing areas of the human mind/brain are not entirely independent, but it is uncertain whether they influence each other because they are nearby, such that activation of one "primes" activation of the other, or because they overlap, such that development of certain musical processing areas would simultaneously develop the particular type of spatial reasoning defined as spatial-temporal.

Further research needs to disentangle the cognitive mechanism that causes the effect. For example, neither priming model--either Shaw's "trion" or Parson's "rhythm" models--is conclusively affirmed or refuted, although both remain promising. In addition, future research needs to distinguish the effect conclusively from potential artifacts of procedures (e.g., subjects' attention to musical stimuli, or subject or experimenter effects that align results with unconscious expectations of subjects or researchers) or research design variations (e.g., control stimuli that are equally preferred by subjects or that can be measured as equally arousing or mood-altering).

The analysis does not have direct implications for education, since the experiments were not about learning, but rather about how the human mind processes two types of information, musical and spatial. However, the result does suggest that studies in which subjects are taught music could plausibly result in spatial learning. A group of such studies were synthesized by Hetland (2000b), and are described below.

The lack of mechanism for the Mozart Effect finding means that the effect is still questionable, and future research may yet demonstrate that the effect is an artifact of research design. While the best evidence to date is that the effect appears to hold up, that does not imply that policy should mandate listening to classical music for any audience. Future research needs to test specific hypotheses about the mechanism underlying this effect, but this laboratory finding with college students implies nothing for the education of children, much less infants *in utero*. If parents or teachers wish to play classical music for themselves or their children, they should by all means do so for any number of reasons. But based on what we know at present, no one should expect that listening to music alone will aid children's future scores on standardized tests of academic achievement.

Music Instruction and Spatial Reasoning. A second body of studies has often been confused with the 'Mozart effect' studies, but it deserves consideration in its own right. Hetland (2000b) identified 19 studies in which children ages 3-12 engaged in programs of active music instruction for up to two years.[4] The studies included in the music instruction analysis were conducted in schools or other instructional settings and used a variety of musical pedagogies and measures of spatial reasoning. To be included in the analysis, studies had to have one or more control conditions, with or without an alternate treatment. About one-third had an alternate treatment for controls consisting of instruction in language, instruction in reading or mathematics

on the computer, or passive instruction in music. Almost all had a non-treatment control (17/19), either in addition to a treated control group or as the only comparison group.

In these studies, music instruction involved combinations of the following: singing, playing musical games, learning notations, improvising or composing music, moving responsively to music, including clapping, and playing instruments. The instruments used in the programs were combinations of voice, piano, xylophones, snare drum, and classroom rhythm instruments (triangles, tambourines, rhythm sticks, finger cymbals, hand-chimes, and bells).

Measures used in the studies varied widely, and because the results of the Mozart Effect analysis indicated that only spatial tasks defined as spatial-temporal were enhanced by music, type of task became the distinguishing feature for three groups of studies analyzed in separate meta-analyses. The first analysis included studies that employed spatial-temporal tasks, the second included studies employing nonspatial-temporal tasks, and the third included studies that employed a variety of spatial tasks that could not be clearly distinguished by the criteria for spatial-temporal tasks.

The first instructional analysis included 15 studies (701 subjects) employing such spatial-temporal tasks as the Object Assembly subtest from the WPPSI-R or WISC-III, in which children assemble a puzzle of a familiar object without seeing a model of the completed image (See Figure 4 for a sample item). Studies using other tasks were also included: a program designed by Matthew Peterson in Gordon Shaw's lab used a measure called the Spatial-Temporal Animation Reasoning or STAR, and other studies used spatial subtests of other standardized tests for children (i.e., Developing Cognitive Abilities Test, the Wide Range Assessment of Visual Motor Abilities, and the Kaufman, Woodcock-Johnson, and McCarthy batteries).

The average effect size was large by meta-analytic standards (r = .37, d = .79), and the results were highly generalizable (t test of the mean Zr was 7.50, $p < .0001$). Most interestingly, despite great variation in the music programs and spatial-temporal measures employed, there was relatively little variation in effect size among the studies included. All had effects greater than zero, the 95% confidence interval was r = .26 to r = .48, the *SD* was less than half the size of the effect at .16, and studies were decidedly drawn from a single population (χ^2 (14) = 20.37, p = .12). We can conclude from these results that the analysis is highly robust.

Contrast analysis of 17 potential moderator variables explored potential reasons for the effect found in this analysis. The most interesting finding is that 13 of these moderators did not influence the size of the effect systematically, even though many of them are factors that often have been found to influence learning. These potential moderators include socio-economic status, duration of instruction, parental involvement, test reliability, teacher and experimenter expectancy effects (unconscious expectations of subjects or experimenters that bias results), the Hawthorne effect (a tendency of any new program to have a positive impact), methods of group assignment, and study quality. In addition, and of particular interest to music educators, keyboard instruction proved no more influential than the other forms of active music instruction tested, despite a reasonable assumption that the spatial layout of the keyboard might be an important contributor in enhancing spatial outcomes. In addition, effect sizes did not vary for those studies that used different keyboard instruments (pianos and xylophones), nor for studies that either did or did not use responsive movement in the music program, nor for studies that either did or did not ask students to create or improvise musically. In other words, the large effect found for the analysis is very stable in relation to a host of variables that might have affected it one way or the other. The effect is not an artifact.

There were, however, two moderator variables that did impact the size of effect. Effect sizes were somewhat larger in studies with individual rather than group lessons, and in studies in which children learned standard notation (rather than either no notation or preparatory types of notation such as Kodaly hand signs). However, the more relevant finding from a policy perspective is that large effects were obtained in both group and individual formats (group lessons $r = .32$, individual lessons $r = .48$) and with and without standard notation (no notation: $r = .36$, standard notation: $r = .39$).

There were also two moderator variables that were nearly significant. The first is the publication status of the article (published articles $r = .29$, unpublished articles $r = .47$). Publication status is often used as a proxy to index study quality, however, a direct analysis of quality showed no difference between studies with higher and lower ratings on threats to internal validity, so the publication result has not been adequately explained. The other variable of interest was subject age (comparing 3-5 years olds to children 6 years of age or older). Since the comparative effect sizes of the two groups were fairly large (3-5 years, $r = .44$, $\geq$ 6 years $r = .27$), the effect is noteworthy. Future research should test whether enhancing effects from music programs are greater for younger children, as was the case here.

The second instructional analysis (5 studies, 694 subjects) included studies with Raven's Matrices as the outcome measure. Based on the results of the contrast on measures in the Mozart Effect analysis, which found a lower average effect ($r = .04$) for nonspatial-temporal measures compared to spatial-temporal measures ($r = .20$), a lower effect size could be anticipated for this analysis. That proved to be the case. The average effect for the nonspatial-temporal measures analysis in the instruction studies ($r = .08$, $d = .16$) was much lower than the average effect of the spatial-temporal measures analysis ($r = .37$, $d = .79$). The average weighted r was even lower (r

= .03, d = .07), which may be the more informative statistic, since four of the studies were similar in size (ranging from 147-179 subjects) and only one differed (40 subjects). The effect was not generalizable (the 95% confidence interval spans zero at r = - .10 to r = .27, t test of the mean Zr = 1.23, p = .29), and the studies were from a single population (χ^2 (4) = 5.72, p = .22). This result provides support for the claim that the effect of music instruction is specific to spatial-temporal and not non-verbal tasks generally, such as Raven's, that rely more on general logic.

The third instructional analysis included 9 studies (655 subjects) that employed a range of spatial measures not readily classifiable as either spatial-temporal or nonspatial-temporal. Thus, this analysis tested whether the enhancing effects of music instruction extend beyond spatial-temporal measures to other, less clearly defined, types of spatial reasoning. Some studies used both spatial-temporal and nonspatial-temporal measures (i.e., several used more than one spatial subtest from the WPPSI-R and only reported a global score), some used tests that may be spatial-temporal but that are difficult to classify (e.g., Children's Embedded Figures Test, or "drawings and words presented in lacunary and ambiguous form" Zulauf, 1993/1994, p. 114). One study used a task that relies mainly on spatial memory (Bead Memory task from the Stanford Binet: Fourth Edition).

The average effect found in this analysis (r = .26, d = .55) is lower than the effect in the spatial-temporal analysis, but it is still of moderate size. In addition, it is generalizable (95% confidence interval r = .16 to r = .36; t test of mean Zr = 6.11, p = .0003), and represents a single population of studies (χ^2 (4) = 8.87, p = .35). From this we can conclude that music instruction may not be limited to spatial-temporal tasks but may enhance spatial reasoning more broadly. Further research is needed to affirm this finding, however, since the measures are quite diverse.

For the instructional analysis, there is a solid, generalizable finding that, for children aged 3-12, active instruction in music—not listening alone, although listening is a component of such instruction—enhances performance on a specific type of spatial task classified as "spatial-temporal." Further, the third instruction analysis for mixed spatial measures suggests that this enhancement may extend more broadly to some nonspatial-temporal forms of reasoning, although not to matrices tasks (as shown in the second analysis).

However, before policy-makers mandate music instruction as a means to enhance children's spatial abilities, important questions about the value to education of such an effect need to be raised. Remember that not all types of music programs have been tested, and that, in fact, the musical treatments combined may be different from each other in important and as yet unspecified ways. More research describing the components of music instruction is needed to clarify just what teachers and students do in music instruction that aids skill in spatial reasoning. Further, the music studies analyzed were only for students between ages 3 and 12, so we cannot generalize to infants, toddlers, or adolescents. Further, because the spatial tests were conducted within a few weeks of the end of the music instruction, we do not know how long any enhancing effect lasts. And because only one longitudinal study extending beyond two years currently exists, and that showed students without music instruction catching up to those with piano instruction during the third year of instruction (Costa-Giomi, 1999), we do not know if music instruction is effective in fostering spatial reasoning after the first two years of instruction.

Perhaps even more important is the question of whether the effects of music instruction on spatial tests translate to better success in school. They might, or they might not. First, "real world" spatial problems, whether found in mathematics or the block corner or the ball field, may or may not be predicted by success on paper and pencil or table-task tests such as those used in

these studies. Second, a corollary to this problem is that many classrooms do not give students a chance to use spatial skills, because instruction may not offer opportunities to apply spatial reasoning to school subjects. In such cases, unfortunately, enhanced spatial ability would not necessarily lead to improved success in school. To reap the benefits of any enhancement of spatial reasoning resulting from music instruction, therefore, schools would also need to insure that instruction emphasizes spatial approaches to learning. Third, because spatial reasoning is multi-dimensional (consider the differences in designing a bridge, packing a car trunk, or finding your way around a new city, for example), it is not clear where the effects of the specifically "spatial-temporal" tasks would show up. Thus, although this is a solid finding, its implications for educational policy are not self-evident.

No Support for Five Instrumental Claims

Arts Rich Education and Verbal and Mathematical Achievement. Perhaps the most commonly heard instrumental claim for the arts is that they lead to enhanced standardized test scores, higher grades, and lowered high school drop out rates. Just what is the evidence for such claims?

Winner and Cooper (2000) synthesized studies that examined the relationship between studying the arts (type of art course was not specified) and verbal and mathematical achievement. These studies do not allow us to determine which form or forms of arts students studied. Thus, all we can say about this body of data is that it examines the effects of studying the arts (which could mean intensive study of some combination of visual arts, music, drama, and dance) on academic achievement. Because our meta-analyses combine studies that examine the effects of a variety of art forms, we refer to these as "multi-arts" meta-analyses.

In the studies synthesized, students were either exposed to the arts as separate disciplines, or they received such exposure but were also given an arts-integrated academic curriculum. Unfortunately, few of the studies explained in much detail anything about the nature and quality of the arts instruction, or about what it really meant to study an academic subject with arts integration. Academic achievement in these studies was measured primarily in the form of test scores (composite verbal and quantitative scores, or verbal and quantitative scores separated) but also sometimes in the form of academic grade point averages or receipt of academic awards.

We first examined the correlational studies—studies that compared the academic profile of students who do and do not study the arts either in school or in after school programs. For example, we included in the analysis James Catterall's study in which he demonstrated that students who are highly involved in the arts in middle and high school outperform those who are not involved in the arts on a multitude of academic indicators, and this relationship holds even for students in the lowest SES quartile of the United States (Catterall, 1998; Catterall, Chapleau, & Iwanaga, 1999). These students earned higher grades and test scores than those not arts-involved. The high arts students were also less likely to drop out of high school and they watched fewer hours of television than did the low arts students. We included Shirley Brice Heath's (1998) study showing that at risk students who participate in after-school arts organizations for at least nine hours a week over the course of at least a year are ahead of a random national sample of students on a wide range of academic indicators: their school attendance is higher, they read more, and they win more academic awards. And we included data from the college board revealing that the average SAT scores of students with four years of high school arts was higher than the scores of those who took no arts courses at all in high school (College Board, 1987-1997).

Three meta-analyses synthesizing the correlational studies were performed, each on a different academic outcome (composite verbal and quantitative outcomes summed; verbal outcomes; quantitative outcomes). All three correlational analyses showed a clear relationship between academic achievement and studying the arts. All three effect sizes were significantly different from zero, as shown by a *t* test. When we examined the five studies that used composite outcomes (verbal and mathematics achievement indicators summed), we found a small but highly significant relationship (r=.05, equivalent to d = .10, 95% confidence interval r = .03 to r = .08, t test of the mean Zr = 5.97, p = .004). When we examined the eleven studies that used verbal outcomes (and this included ten years of the College Board data), we found a small to medium relationship (r=.19, equivalent to d = .39) which was also highly significant (95% confidence interval r = .17 to r =.22, t test of the mean Zr = 16.52, p <.0001). And when we examined the eleven studies that used mathematics outcomes (and this included ten years of the College Board data), we again found a small to medium relationship (r=.10, equivalent to d = .20) that was highly significant (95% confidence interval r = .07 to r =.14, t test of the mean Zr = 6.36, p <.0001).

These three meta-analyses show that students in the United States who choose to study the arts are students who are also high academic achievers. But because the studies on which these meta-analyses were based were correlational in design, they allow no causal inferences. Does art study cause higher scores? Or do those with higher scores take more art? Or, is there a third variable, such as parental involvement, that causes both greater arts study and higher test scores? We cannot tell. Unfortunately, however, studies such as these have often been used erroneously to support the claim that studying the arts *causes* test scores to rise.

One plausible non-causal interpretation of the findings is that high academic achievers (no matter what their SES) may be more likely to choose to study the arts than low academic achievers. This could occur for several reasons. High academic achievers may attend schools strong in both academics and the arts; they may come from families that value both academics and the arts; or they may have high energy and thus have time for and interest in both academics and the arts.

One piece of evidence for the high energy hypothesis comes from the study by Heath (1998). Heath's study included not only students involved in after-school arts organizations, but also those in two other kinds of after-school organizations, those focussing on sports, and those focussing on community service. All three groups were intensively involved in their choice of organization. Heath allowed us access to her unpublished data, and we compared the likelihood of winning an academic award for the arts vs. the sports students. While both groups were significantly more likely to win an academic award than a random national sample of students, we found no difference between these two groups. Eighty-three percent of the group of 143 arts-involved students and 81% of the sports-involved students won an academic award, compared to 64% of the national sample. The finding that both intensively involved sports and arts students did well academically is consistent with (though does not prove) the possibility that these are highly motivated students to begin with. Perhaps the drive factor is what impels these students both to involve themselves in an after school activity in a serious way as well as to do well in school. It is also possible that these students get "hooked," whether on sports or arts, and when they are thus engaged their energy is productively channeled.

Some support for the drive hypothesis comes from a comparison pointed out by Eisner (2001). He compared the SAT advantage of students taking four vs. one year of arts to that of

students taking four vs. one year of an elective academic subject such as science or a foreign language. Students who specialized in any subject, whether arts or an academic elective, all had higher SATs than those who had only one year in that subject (with academic specialization yielding a far greater advantage than arts specialization). For example, in 1998, while students with four years of arts had verbal SAT scores that were 40 points higher than those with only one year of arts, those with four years of a foreign language had verbal SAT scores that were 121 points higher than those with only one year of foreign language. Similarly, while students with four years of arts had mathematics SAT scores that were 23 points higher than those with only one year of arts, those with four years of science had mathematics SAT scores that were 57 points higher than those with only one year of science. Students who specialize or focus might have higher energy than those who do not, and this higher drive could account for their higher academic achievement. It is also possible, however, that the very process of sticking to something (whether art or an academic subject) leads to better academic performance in other areas.

Another reason for the strong correlation found between arts study and SAT scores could be that our highest achievers study the arts in order to enhance their chances of admission to selective colleges. It should be noted, in this regard, that the academic profile of students choosing to take the arts has risen consistently over the last decade. When Vaughn and Winner (2000) plotted the relationship between SAT score and taking four years of arts in high school (compared to taking no arts), we found that this relationship grew stronger each year beginning with the first year in which the data are available (1988) and continuing through 1999 (the last year of data we examined). Rising effect sizes for the arts-SAT relationship are shown in Figure 5. Thus, the comparative SAT advantage for students with four years of arts grew greater each

year. As our most selective colleges become more competitive each year, students may feel they need to build resumes showing strength in a non-academic area such as an art form.

An examination of the relationship between arts study and academic achievement in other countries proves extremely instructive. In the Netherlands, Haanstra (2000) found that students who take the arts in high school to prepare for a national exam that includes the arts attain the same educational level as those with no arts electives. This study, which controlled for students' SES, shows that in the Netherlands, taking the arts in high school does not predict ultimate educational level attained. In the UK, Harland and colleagues (Harland, Kinder, Haynes, & Schagen, 1998) found that the greater the percentage of arts courses taken in high school, the poorer the performance on national exams at the end of secondary school. Harland explained this finding by noting that in the UK, the only students who are permitted to prepare for more than one arts subject for their secondary school exams are those who are academically weak. This contrasts sharply with educational policy in the United States. Academically weak students in the US are steered into remedial academic courses, not into the arts. The comparison between the findings in the United States with those in the Netherlands and the UK suggest that the relationship between arts study and academic achievement is not a causal one but instead reflects different cultural values about who should study the arts.

We reasoned that even if self-selection (high achievers choosing to study arts) explains the correlation in the US, there might still be some causal force at work. Might it not be that once high achievers self-select into the arts, the arts then foster cognitive skills which translate into even higher academic performance? We were able to test this hypothesis by examining the data in James Catterall's study mentioned earlier (Catterall, 1998; Catterall, Chapleau, & Iwanaga, 1999). Catterall reported longitudinal data on students who self-selected into the arts in 8th grade

and remained highly involved in the arts through the 12th grade. If both factors were at work, we would expect the effect sizes showing the strength of the relationship between arts involvement and academic performance to rise over the years. But we found no change. The effect size showing the relationship between studying the arts and academic achievement was *r*=.18 (equivalent to *d* = .37) for students in 8th grade, and this effect size remained unchanged in 10th and 12th grade. Although these data come from only one study, they come from a very large-scale study: there were 3,720 students who were highly involved in the arts from the 8th through 12th grades, and the same number who were not particularly involved in the arts over that time period. The data fail to support the view that the arts are what is causing the academic achievement of these students to be higher than that of students relatively uninvolved in the arts.

While the correlational studies, and the meta-analyses synthesizing them, do not permit causal inferences, studies with an experimental design do allow such inferences. We examined two bodies of experimental studies testing the causal claim that when students study the arts, their academic achievement rises. These studies compared academic performance before and after studying the arts. Typically these studies examined students at the elementary school level who had studied the arts for a year and who studied the arts both as separate disciplines and as integrated into the academic curriculum. The academic growth of these students was then compared to the growth of similar students not exposed to any special arts program.

We found 24 studies testing the hypothesis that verbal skills improve as a consequence of studying the arts, and 15 studies testing the hypothesis that mathematics skills improve. The meta-analysis performed on the verbal outcomes yielded a mean effect size *r* of .07 (equivalent to *d* = .14). This effect size was not statistically significant. The 95% confidence interval was *r* = .01 to *r* =.14. In addition, a *t* test of the mean *Zr* showed that the mean effect size found was not

significantly different from zero. Moreover, the 19 studies in which the arts were integrated into the curriculum yielded a mean effect size identical to that of the five studies in which the arts were only studied separately. Thus we had to conclude that we had found no evidence that studying the arts, including the arts integrated with academic subjects, resulted in enhanced verbal skills.

The meta-analysis performed on the mathematics outcomes yielded a mean effect size of r=.06 (equivalent to d = .12). Again the 95% confidence interval included zero, and the □□test of the mean *Zr* showed that the mean effect size was not significantly different from zero. In this case we could not statistically compare the studies with and without arts integration since all but two were based on an arts integrated curriculum. Again, then, we had to conclude that we found no evidence that studying the arts, including the arts integrated with academic subjects, resulted in enhanced mathematics achievement.

Thus we can see that there is (yet) no evidence that studying the arts, or studying an academic curriculum in which the arts are somehow integrated, results in higher verbal and mathematics achievement, at least as measured by test scores, grades, or winning academic awards.

Arts Rich Education and Creativity. Does studying the arts lead to enhanced critical and creative thinking outside of the arts? This claim seems more plausible than the claim that the arts lead to higher verbal and mathematical test scores, and we felt optimistic about this section of our research. Unfortunately, we found no studies testing this claim by assessing any kinds of thinking skills besides those measured by standard paper and pencil creativity tests (Moga, Burger, Hetland, & Winner, 2000). We found four studies comparing the creativity test scores of students who took arts courses vs. those who did not. When we entered the verbal creativity

scores into a meta-analysis, we found r = .05, equivalent to d = .10. This relationship was not statistically significant at p=.64 (95% confidence interval r = -.21 to r = -.31, t test of the mean Zr = .81, p = .50). We did find a small to medium sized relationship (r=.19, equivalent to d = .39) between studying arts and figural creativity tests (which themselves are visual tests) but even this relationship did not withstand the most important significance test since it was not significantly different from zero (95% confidence interval r = -.05 to r =.44, test of the mean Zr = 3.19, p = .09). It seems reasonable to suggest that paper and pencil creativity tests are not the right kinds of outcomes to be using, as these tests primarily assess fluency and cleverness. Future research should examine more qualitative creative thinking outcomes, such as the ability to find new problems (Getzels & Csikszentmihalyi, 1976).

Visual Arts and Reading. Can studying the visual arts help remedial readers improve their reading? This is the assumption guiding several programs set up in New York City, such as the Guggenheim Museum's Learning to Read through the Arts, Reading Improvement Through the Arts, and Children's Art Carnival. In these programs, children with reading difficulties are given experience in the visual arts which is integrated with reading and writing. For example, children drew and then wrote and read in connection with what they drew. These programs generally find that remedial readers improve their reading scores quite considerably. They then conclude that this improvement is due to the arts experience students received. Unfortunately, these programs failed to compare the effects of an arts-reading integrated program with the effects of an arts-alone program. Therefore we cannot know whether the reading improvement that undoubtedly did occur was a function of art experience, art experience integrated with reading, or simply from the extra reading experience and instruction.

We examined two groups of studies: those that compared an arts-only instruction to a control group receiving no special arts instruction (nine studies); and those that compared an art-reading integration treatment to a control group receiving reading only (four studies). The first group allowed us to see whether instruction in visual art by itself teaches skills that transfer to reading skills; the second group allowed us to test whether reading integrated with art is more effective than reading instruction alone.

A meta-analysis of the studies testing the effects on reading of art instruction alone yielded a small effect (r=.05, equivalent to d = .10) which could not be generalized to new studies (95% confidence interval r = -.30 to r =.54, t test of the mean Zr = .53, p = .61). A meta-analysis of the studies testing the effects of art-reading integrated instruction yielded a mean effect size of r = .23 (equivalent to d = .47), and again this result could not be generalized to new studies (95% confidence interval r = .03 to r =.45, t test of the mean Zr = 2.003, p = .14). Moreover, this effect was entirely due to reading readiness outcomes, and these are visual outcomes. There was no effect for reading achievement outcomes.

Thus we had to conclude that there is no support for the claim that the visual arts enhance reading skills. Programs that help remedial readers improve their reading through a reading-arts integrated program are likely to work well because of the extra intensive reading training that the children receive, independently of the fact that this training is fused with drawing.

Dance and Reading. It is difficult to imagine how dance could enhance reading at the level of decoding, though one could hypothesize that by enacting stories through dance, comprehension of these stories might deepen. In Chicago, a program called Whirlwind had sought to improve basic reading skills in young children through dance (Rose, 1999). One of the activities that children in this program engage in is "dancing" their bodies into the shapes of

letters. By virtue of this activity, these children in fact improved their beginning reading skills significantly more than did a control group which did not get the same kind of "dance" instruction. Unfortunately, however, we cannot conclude that the dance activity is what led to the reading improvement since the control group did not get the same kind of letter training. It must be added, as well, that the activity of putting one's body into the shape of letters is not authentic dance, though in fact it may prove to be an excellent way of helping children remember letters.

We searched for studies that examined the effect of dance on reading which also had appropriate control groups (Keinanen, Hetland, & Winner, 2000). A meta-analysis on the four identified studies showed a small effect size between dance and reading (r=.10, equivalent to d = .20), but this effect size was not significantly different from zero (95% confidence interval r = -.21 to r =.42, t test of the mean Zr = 1.03, p = .38). Thus, we concluded that there is no evidence that dance is a tool to enhance reading. However, the main finding of this analysis is the paucity of studies that test the relationship between dance and non-arts learning of any kind. Until more studies are conducted, the case cannot be made convincingly one way or the other.

Music and Reading. Music has also been claimed to be a way to improve reading skills, possibly because of the effect of learning to read music notation. In reading of both text and music notation, the written code maps onto a specific sound; hence, perhaps practice in reading music notation paves the way for learning to read linguistic notation. In addition, perhaps listening to music trains the kind of auditory discrimination skills needed to make phonological distinctions. It is also possible that music enhances reading skills only when students learn to read the lyrics of songs.

As part of the REAP project, Butzlaff (2000) located six experimental studies testing music's effect on reading and performed a meta-analysis on these studies. He found a mean

effect size of *r*=.18 (equivalent to *d* = .37). This average was based on quite varied effect sizes, and the effect size was not significantly different from zero (95% confidence interval *r* = -.21 to *r* = +.52, *t* test of the mean *Zr* = 1.06, *p* = .34). So, we have to conclude that there is no evidence thus far that learning music aids the development of reading.

Equivocal Support for Two Instrumental Claims

Dance and Spatial Reasoning. Keinanen et al. (2000) were able to find four studies assessing the effect of dance instruction on nonverbal, performance IQ scales and on nonverbal paper and pencil spatial reasoning tests. The average effect size yielded by a meta-analysis on these studies was *r*=.17 (equivalent to *d* = .35), and this was statistically significant (95% confidence interval *r* = .06 to *r* =.29, *t* test of the mean *Zr* = 3.46, *p* = .04). We can conclude that dance does enhance nonverbal skills. This finding constitutes a case of near transfer and is not surprising since dance itself is a visual-spatial form of activity. In addition, although it is a positive relationship, it is based on very few studies. The bigger story in dance remains that very little research has been conducted to test rigorously the relationship between dance and non-arts learning.

Music and Mathematics. In 1999, a study published in *Neurological Research* received a lot of publicity (Graziano, Peterson, & Shaw, 1999). This study reported that piano keyboard training along with computer-based spatial training led to greater improvements in mathematics than when spatial training was combined with computer-based English language training. Vaughn (2000) searched for other studies examining the power of music to stimulate mathematical thinking and found six. Meta-analysis of these studies found an average effect size of *r*=.13 (equivalent to *d* = .26), the confidence interval did not span zero (95% confidence interval *r* = .03 to *r* =.23) and the *t* test of the mean *Zr* = 2.49, which was nearly significant

(considering the .05 level as a cut-off) at p =.06. These findings suggest that there may indeed be a causal link between some forms of music instruction and some forms of mathematics outcomes. But no firm conclusions can be drawn at this point since the finding was based on only six experiments. Moreover, of these six results, only two yielded medium sized effects (r=.31, .20, equivalent to d = .65, .41), one yielded a small to medium sized effect (r=.17, equivalent to d = .35), and the remaining three were below .10, the level considered to be small (one of which was actually negative.) Thus, more research on this question is needed before we can be sure about the result.

Research Implications of REAP

> Although the findings are not entirely negative, and although the limits of the analysis are carefully articulated by the authors, it is important to stand back from their findings and ask whether the game is essentially over…. Some would say that it had never really begun (Perkins, 2001, p. 117).

Meta-analytic syntheses such as those conducted during REAP are not the final word on a research area. Instead they clarify what the research has thus far shown and guide attention to questions that remain to be asked. The REAP research summarized here assesses what we know to date about cognitive transfer from arts education to non-arts learning. In addition to informing policy-makers about what research has to tell us about transfer from arts to non-arts learning at this point in time, the results of the REAP analyses can be used to guide future studies on this complex question. And, as David Perkins suggests in the above quotation, one of the implications of REAP is that, as arts education researchers, we need to play a better game about transfer from arts to non-arts learning.

A better game, in our view, means that we need to (1) shift the areas of research focus and (2) refine the research methods.

Shifting the Focus

First, we believe that the field needs a renewed focus on teaching and learning *in* the arts. To continue building strong practice and to provide support for doing so, both policy-makers and practitioners need descriptions and evidence for what arts instruction achieves at its core. It is the responsibility of researchers to provide that evidence.

Second, we need research that examines possible non-cognitive transfer outcomes of arts education: the social, motivational, or dispositional effects of arts instruction. For example, when schools take the arts seriously, do they become more inclusive environments, more tolerant of differences, more focused on social justice? Do students in such schools attend more regularly, stay in school longer, work in a more disciplined manner in non-arts subjects, and/or show a willingness to reflect on and revise their work in non-arts subjects?

Third, we need to investigate how other subject areas can learn about good teaching and deep learning by looking at arts classes. For example, we might test the effects of arts-as-entry-points in a variety of subjects: Do students with certain kinds of profiles engage more deeply with subject matter when arts are used as entry points? Which students? How? And when? Would students in mathematics or English classes benefit from greater proportions of class-time being devoted to working on projects while teachers offer individual consultations of ongoing work, similar to the way studio art courses are run? Or would science, history, or language classes benefit from the kind of regular, mid-project critiques that are common in studio arts courses?

Fourth, we need to search for reasonable "bridges" between specific arts and specific subject matters. It may be more reasonable to expect transfer from the arts to higher-order cognition (reflection, critical thinking, creative thinking, ability to tolerate ambiguity and resist premature closure when solving a "messy problem" with no clear right answer) than to more basic level skills such as spelling or vocabulary (Eisner, 2001; Perkins, 2001; Tishman, MacGillivray, & Palmer, 1999).

Fifth, we need to examine the effects of *explicit* teaching for transfer in the arts. Perhaps it is only when teachers make clear that the skills being taught in arts classes can be used in other subject areas, can help students see how they might do so, and/or can work with students to reflect on and practice making such connections that students become able to transfer skills learned in the arts.

Improved Research Methodology

Research in arts education also needs to be improved methodologically. Our reviews of the literature revealed that many researchers in the arts have not applied standard social science methodology for the rigorous conduct and reporting of research. This may be due to the fact that much arts education research has been conducted in schools, rather than in the laboratory, and field-based research is always more complex than laboratory research. In addition, arts education research has been hampered by not having been routinely well-funded; hence, arts researchers have not had as many opportunities to learn from mistakes as have researchers in mathematics or science or reading. But we should face the need for improved methodology, rather than feel defensive. Doing so is what will help advance understanding of the complex issues surrounding teaching and learning in the arts.

Perhaps the first and most important research implication from REAP is that the field needs to embrace the value of synthesis: more arts education scholars need to develop skill in meta-analysis (cf. Rosenthal & Hetland, 2001). Because meta-analytic procedures and methods are codified and described explicitly, meta-analytic reviews are more replicable than traditional reviews and allow reviewer bias to be revealed over time through the scientific process. In this way, the weaknesses of meta-analysis, which in our view is the best (though imperfect) way to summarize research, will improve. With improved synthesis, our findings become more trustworthy.

A corollary to the need to embrace synthesis is the need for clearer reporting of empirical research. Shouldn't arts journals require that studies report effect sizes, exact quantities for all significance tests performed (i.e., *t, F*, or X^2), their associated degrees of freedom and exact *p* levels (even for "non-significant" findings),[5] confidence intervals, tables of *N*s, means, and standard deviations for all groups, and ANOVA or regression tables where appropriate? All of these quantities are necessary for accurate synthesis, and accurate synthesis is necessary if practice and policy are to rely on research. In addition to such reporting, researchers need to identify threats to validity and alternative explanations of results. These always exist and always require explanation. And finally, any study of arts learning, whether learning in the arts or learning transferred from the arts, should report clear descriptions of teaching methods, since the characteristics and quality of teaching certainly affect how well students can use what they learn flexibly and appropriately.

A second research implication from REAP is that we need to end the pointless debate about whether qualitative or quantitative paradigms are most appropriate in the arts. Both methodologies aid our understanding of the complex and subtle phenomena involved in artistic

learning and practice. Skill with one paradigm supports skill in the other. Quantitative research is not inherently reductive; qualitative research is not inherently fuzzy. Researchers in the arts need to be trained in both qualitative and quantitative paradigms and then employ the methods that best suit the questions they wish to answer.

A corollary to the appropriate use of paradigms is that arts education researchers should learn from innovative methods developed in other disciplines. Many areas of education (e.g., mathematics, reading, writing, science), and other social, behavioral, and biomedical areas of research, including social and clinical psychology, anthropology, sociology, medicine, and public health, have had more resources devoted to research over time than have the arts. The arts would be wise to use the good fortune of these domains as sources of information about how sophisticated methodologies can help us answer questions of interest in our own field.

A third implication is that we need to design studies more rigorously. "You can't fix by analysis what you bungle by design" (Light, Singer, & Willett 1990). Such rigor includes conducting two kinds of studies: those that develop theories and those driven by theory, with both kinds focused on defining the mechanisms that link treatments and outcomes. Longitudinal designs need to be conducted more commonly, and contrasts need to be employed more routinely. Assignment to experimental groups should be randomized at the level of the individual whenever possible and, when not feasible, matched at least for IQ, SES, parental education, and parental arts background. And studies must employ control groups with alternative treatments (besides arts) so that specific hypotheses can be tested and potential confounds disentangled.

A fourth research implication from REAP is that we need to turn our attention to the development of measures. If we value students learning, for example, to perceive, think, and understand in addition to acquiring technique and memorizing information, then we need to

develop tests that allow students to demonstrate, and teachers, states, and researchers to assess, those qualities. An under-utilized technique in arts assessment is rating by expert judges. It is central in the arts (e.g., in the assessment of portfolios for admission to arts schools, in qualifying processes for juried exhibitions) and in other disciplines in which nuance separates levels of quality (e.g., judging of figure skating and gymnastics at the Olympic level). And, when cognitive transfer from the arts to other subjects is of interest, researchers need to include measures of learning in the art form itself and compare that to learning in the other subject(s). Higher levels of transfer outside of the arts should reflect greater learning in the parent domain (Bransford & Schwartz, 1999).

With attention to these topics and methods, arts education research will be able to advance quickly from the benchmark defined by REAP in 2000.

Policy Implications of REAP: How Should We Justify Arts Education?

Perhaps the most important policy implication of the research reported here is that arts education policy should not be based on instrumental outcomes for the arts, whether or not these outcomes can be demonstrated. If they cannot be demonstrated, the case is clear: we must make honest arguments for the importance of the arts. But even in cases where they can be demonstrated, we should not use instrumental outcomes as justifications. We need to distinguish between core justifications for teaching the arts versus instrumental ones. Core justifications are the central reasons: they are about learning in the disciplines of the arts themselves. Instrumental reasons are the side effects–enhanced learning in non-arts disciplines, which may or may not occur. It is self-destructive to justify the arts on the basis of instrumental effects. If the arts are given a role in our schools because people believe the arts cause academic improvement, then the arts will quickly lose their position if academic improvement does not result, or if the arts are

shown to be less effective than direct instruction in literacy and numeracy. Instrumental claims for the arts are a double-edged sword. It is implausible to suppose that the arts can be as effective a means of teaching an academic subject as is direct teaching of that subject.

When instrumental reasons become the chief justification for arts education, arts teachers may feel compelled to teach the arts in a way that will enhance academic (rather than artistic) understanding. They may turn strings of music notations into multiplication problems and bill this as music education, the kind likely to improve mathematics scores. Or they may teach the physics of sound in music class rather than the aesthetics of sound, or have students build musical instruments (because that may improve their spatial abilities) rather than learn to play these instruments.

It is time to state the right arguments for the arts in our schools and to begin to gather the right kind of evidence for these arguments. The best hope for the arts in our schools is to justify them by what the arts can do that other subjects cannot do as well, or cannot do at all.

The two most important reasons for studying the arts are to enable our children to be able to appreciate some of the greatest feats humans have ever achieved (e.g., a Rembrandt painting, a Shakespeare play, a dance choreographed by Martha Graham, a Charlie Parker jazz improvisation), and to give our children sufficient skill in an art form so that they can express themselves in this art form. The arts are the only arenas in which deep personal meanings can be recognized and expressed, often in nonverbal form.

In reaction to our work, arts advocates have said that we are just returning to “arts for arts sake” arguments, and that these old arguments just won’t wash. But this is an admission of defeat. If we realize that the arts are as important as the sciences, and that the purpose of education is to teach our children to appreciate the greatest of human creations, then the arts will

have a strong hold in our schools. But if we become swayed by today's testing mentality and come to believe that the arts are important only (or even primarily) because they buttress abilities considered more basic than the arts, we will unwittingly be writing the arts right out of the curriculum.

References

Bransford, J., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24*, 61-100.

Butzlaff, R. (2000). Can music be used to teach reading*? Journal of Aesthetic Education, 34* (3-4), 167-178.

Catterall, J. (1998). Involvement in the arts and success in the secondary school. *Americans for the Arts Monographs 1*(9).

Catterall, J., Chapleau, R., & Iwanaga, J. (1999). Involvement in the arts and human development: General involvement and intensive involvement in music and theater arts. In E. Fiske (Ed.)*, Champions of change: The impact of the arts on learning* (pp. 1-18).\. The Arts Education Partnership and The President's Committee on the Arts and the Humanities.

Chabris, C. (1999). Prelude or requiem for the 'Mozart Effect'*? Nature 402*, 826-827.

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

Colwell, R. (2001). The effects of early music experiences. In E. Winner & L. Hetland (Eds.), *Beyond the soundbite: Arts education and academic outcomes* (pp. 89 - 98). Los Angeles: J. Paul Getty Trust.

College bound seniors: A profile of SAT and achievement test takers. The College Board, Princeton, N.J.: 1987-1992; 1994-1997.

Conard, F. (1992). *The arts in education and a meta-analysis.* Unpublished Doctoral Dissertation, Purdue University.

Cooper, H. & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage.

Costa-Giomi, E. (1999). The effects of three years of piano instruction on children's cognitive development. *Journal of Research in Music Education, 47*(5), 198-212.

Dansky, J. L. Cognitive consequences of sociodramatic play and exploration training for economically disadvantaged preschoolers. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 21*, 47-58.

Eisner, E. (2001). What justifies arts education: What research does not say. In M. McCarthy (Ed.), *Enlightened advocacy: Implications of research for arts education policy practice* (pp. 19-29). The 1999 Charles Fowler Colloquium on Innovation in Arts Education. College Park: University of Maryland.

Fiske, E. (Ed.). (1999). *Champions of change: The impact of the arts on learning.* Washington,D.C.: Arts Education Partnership and President's Committee on the Arts and Humanities.

Getzels, J. & Csikszentmihalyi, M. (1976). *The creative vision: A longitudinal study of problem finding in art*. New York: Wiley.

Graziano, A., Peterson, M., & Shaw, G. (1999). Enhanced learning of proportional math through music training and spatial-temporal training. *Neurologoical Research 21* (2), 139-152.

Gregory, R. J. (1996). *Psychological testing: History, principles, and applications (2nd ed.).* Boston: Allyn and Bacon.

Haanstra, F. (2000). Dutch studies of the effects of arts education on school success. *Studies in Art Education 41*(3), 19-33.

Harland, J., Kinder, K., Lord, Pl., Stott, A., Schagen, I., Haynes, J. (2000). *Arts education in secondary schools: Effects and effectiveness*. York, UK: National Foundation for Educational Research.

Heath, S. (1998). Living the arts through language and learning: A report on community based youth organizations. *Americans for the Arts Monographs*, *2*, no. 7.

Hetland, L. (2000a). Listening to music enhances spatial-temporal reasoning: Evidence for the "Mozart effect." *Journal of Aesthetic Education, 34*(3/4), 105-148.

Hetland, L. (2000b). Learning to make music enhances spatial reasoning. *Journal of Aesthetic Education, 34*(3/4), 179-238.

Kardash, C.A. M., & Wright, L. (1986). Does creative drama benefit elementary school students? A meta-analysis. *Youth Theatre Journal 1*(3), 11-18.

Keinanen, M., Hetland, L., & Winner, E. (2000). Teaching cognitive skill through dance: Evidence for near but not far transfer. *Journal of Aesthetic Education, 34* (3-4), 295-306.

Leng, X., & Shaw, G. L. (1991). Toward a neural theory of higher brain function using music as a window. *Concepts in Neuroscience, 2*(2), 229-258.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education.* Cambridge: Harvard University Press.

Moga, E., Burger, K., Hetland, L., & Winner, E. (2000). Does studying the arts engender creative thinking? Evidence for near but not far transfer. *Journal of Aesthetic Education, 34* (3-4), 91-104.

Mosteller, F. & Colditz, G. A. (1996). Understanding research synthesis (Meta-Analysis). *Annual Review of Public Health, 12*(1), 1-23.

Murfee, E. (1995). Eloquent evidence: Arts at the core of learning. Report by the President's Committee on the Arts and the Humanities.

Parsons, L. M., Martinez, M. J., Delosh, E. L, Halpern, A., & Thaut, M. H. (1999). *Musical and visual priming of visualization and mental rotation tasks: Experiment 1*. Manuscript in preparation, San Antonio: University of Texas.

Perkins, D. (2001). Embracing Babel: The prospects of instrumental uses of the arts for education. In E. Winner & L. Hetland (Eds.), *Beyond the soundbite: Arts education and academic outcomes* (pp. 117-124). Los Angeles: J. Paul Getty Trust.

Podlozny, A. (2000). Strengthening verbal skills through the use of classroom drama: A clear link. *Journal of Aesthetic Education, 34* (3-4), 91-104.

Rauscher, F. H., Bowers, M. K. & Kohlbeck, K. (1999). [Mozart effect and laterality.] Unpublished raw data, University of Wisconsin at Oshkosh.

Rauscher, F.H., & Hayes, L. J. (1999). The effects of music on spatial-temporal task performance: Exploring task validity. Manuscript submitted for publication, University of Wisconsin at Oshkosh.

Rauscher, F. H., & Ribar, R. J. (1999). Music and spatial-temporal task performance: Effects of arousal and preference. Manuscript in preparation, University of Wisconsin at Oshkosh.

Rauscher, F. H., & Shaw, G. L. (1998). Key components of the Mozart effect. *Perceptual and Motor Skills, 86*, 835-841.

Rauscher, F. H., Shaw, G. L., & Ky, K. N. (1993). Music and spatial task performance. *Nature*, *365*, no 6447: 611.

Rauscher, F. H., Shaw, G. L., & Ky, K. N. (1995, Feb.). Listening to Mozart enhances spatial-temporal reasoning: Towards a neurophysiological basis. *Neuroscience Letters 185,* 44-47.

Rideout, B. E., Dougherty, S., & Wernert, L. (1998). Effect of music on spatial performance: A test of generality. *Perceptual and Motor Skills, 86,* 512-514 [experiments 1 and 2].

Rideout, F. H., Fairchild, R. A., & Urban, G. E. (1998). *The "Mozart Effect" and skin conductance*. Paper presented at Eastern Psychological Association, Boston, MA.

Rideout, B. E., & Laubach, C. M. (1996). EEG correlates of enhanced spatial performance following exposure to music. *Perceptual and Motor Skills, 82*, 427-432.

Rideout, B. E., & Taylor, J. (1997). Enhanced spatial performance following ten minutes exposure to music: A replication. *Perceptual and Motor Skills, 85*(1), 112-114.

Rose, D. (1999). *The iimpact of Whirlwind Basic Reading through Dance Program on first grade students' basic reading skills: Study II*. Chicago: 3-D Group.

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

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. Hedges (Eds.), *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118* (2), 183-192.

Rosenthal, R. & Hetland, L. (2001). Meta-analysis: Its use and value in arts education research. In E. Winner & L. Hetland (Eds.), *Beyond the Soundbite: Arts Education and Academic Outcomes* (pp. 1-16). Los Angeles: J. Paul Getty Trust.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis.* New York: McGraw-Hill.

Rosenthal, R., Rosnow, R.. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge, UK: Cambridge University Press.

Rosenthal, R., & Rubin, D. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

Salomon, G. & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist. 24*(2), 113-142.

Smilansky, S. (1968). *The effects of sociodramatic play on disadvantaged preschool children* New York: Wiley.

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

Steering Committee of the Physicians Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine, 318*, 262-264.

Thorndike, R. L., Hagen, E. P., Sattler, J. M. (1986). *The Stanford-Binet scale of intelligence*. Riverside, IL: Chicago.

Tishman, S., MacGillivray, D, & Palmer, P. (1999) Investigating the educational impact and potential of the Museum of Modern Art's Visual Thinking Curriculum: Final report. Unpublished manuscript.

Vaughn, K. (2000). Music and mathematics: Modest support for the oft-claimed relationship. *Journal of Aesthetic Education, 34* (3-4), 149-166.

Vaughn, K., & Winner, E. (2000). SAT scores of students who study the arts: What we can and cannot conclude about the association. *Journal of Aesthetic Education, 34,* 3-4, 77 – 90.

Wechsler, D. (1967). *Manual for the Wechsler Preschool and Primary Scale of Intelligence*. New York: The Psychological Corporation.

Wilkinson, L. (1999, August). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604.

Winner. E. & Cooper, M. (2000). Mute those claims: No evidence (yet) for a causal link between arts study and academic achievement. *Journal of Aesthetic Education*, 34(3/4), 11-75.

Winner, E. & Hetland, L. (Eds.). (2000). The arts and academic achievement: What the evidence shows. *Journal of Aesthetic Education, 34*(3/4).

Zulauf, M. (1993/1994). Three-year experiment in extended music teaching in Switzerland: The different effects observed in a group of French-speaking pupils. *Bulletin of the Council of Research in Music Education, 119*, 111-121.

Author Note

Lois Hetland, Project Zero, Harvard Graduate School of Education.

Ellen Winner, Department of Psychology, Boston College, and Project Zero, Harvard Graduate School of Education.

The research reported in this chapter was supported by the Bauman Foundation, and we thank John Landrum Bryant for his support and guidance. Full reports on the meta-analyses summarized here are published in an invited special double issue of the *Journal of Aesthetic Education*, Fall/Winter 2000, Volume 32, Nos. 3-4. We thank Ralph Smith for his editorial guidance. Hetland's music analyses are reported in her dissertation, Harvard Graduate School of Education, 2000. A conference devoted to presentation and critique of the meta-analyses reported here was held at the Getty Center, August 24-25, 2000, and the proceedings of the conference have been published by the J. Paul Getty Trust *as Beyond the soundbite: Arts education and academic outcomes,* 2001.We thank the President and Chief Executive Officer of the J. Paul Getty Trust, Barry Munitz, Deputy Director of the Getty Grant Program, Jack Meyers, and the Getty staff for their support in these endeavors. An Executive Summary of these analyses is published in the *Arts Education Policy Review*, May/June 2001. We thank Sam Hope for his editorial guidance.

We wish to acknowledge the researchers from our project team, whose papers are summarized in this chapter: Kristin Burger, Ron Butzlaff, Monica Cooper, Mia Keinanen, Erik Moga, Ann Podlozny, and Kathryn Vaughn. Additionally, we thank Ron Butzlaff, Robert Rosenthal, and Richard Light for statistical advice throughout the project and Howard Gardner, David N. Perkins, and Judith Singer for generous review and counsel. We also wish to thank the research assistants who helped us at various times throughout the project: Kristin Burger, Lisa

French, Kimberlee Garris, Nandita Ghosh, Maxwell Gomez-Trochez, Jessica Gordon, Joanna Holtzman, Jenny Martin, Elisabeth Moriarty-Ambrozaitis, Brian Moss, Melissa Mueller, Leah Okimoto, Nina Salzman, and Daniel Schneider. Finally, for their generosity and cooperation in answering questions about their work, we thank the researchers whose work we reviewed.

Correspondence concerning this article should be addressed to Lois Hetland, Project Zero, Harvard Graduate School of Education, 124 Mt. Auburn Street, Suite 500, University Place, Cambridge, MA 02138. E-mail: Lois@pz.harvard.edu

Footnotes

[1] Quantitative methods such as contrasts, which were employed extensively in the REAP reviews, explore similar questions through data-analytic procedures, but these methods are possible only in later developmental stages of research when specific directional hypotheses can be advanced based on previous qualitative and quantitative work.

[2] To make this simple calculation, divide the reported effect size *r* by 2 and add it to .50 (e.g., for *r* = .20, 1/2*r* = .10; .10 + .50 = .60 or 60%). That quantity is the percentage of people who are helped by the treatment, so for .60, 60 of every 100, or 600,000 of every million are helped). Subtract that number from 100 for the number not helped (i.e., for this example, 40%, or 400 per 1000, or 400,000 per million not helped).

[3] This is true unless the size of the effect is truly zero, in which case a larger study will not produce a result that is any more significant than a smaller study. Effect sizes of exactly zero, however, are rarely encountered.

[4] Costa-Giomi's study (1999) lasted for 3 years, but only the first two years of data could be analyzed.

[5] The Task Force on Statistical Inference, American Psychological Association, recommends reporting exact *p*s so that distinctions can be assessed along a continuum, rather than at an arbitrarily defined cut off between "true" and "false" When researchers report only whether $p \leq .05$ rather than reporting exact *p*s, "likelihood" is assessed as a cliff. Such reporting of results equates as equally likely probabilities of, for example, $p = .06$ and $p = .50$ (one-tailed), when they are not equivalent. A $p = .06$, one-tailed, indicates that if the null hypothesis were true we would find a t of this size in the predicted direction only 6% of the time. A $p = .50$, one-tailed,

however, indicates that if the null hypothesis were true we would find a t of this size in the predicted direction 50% of the time (Wilkinson, 1999).

Figure Captions

Figure 1: An Item from the Paper Folding and Cutting Subtest of the Stanford-Binet: Fourth Edition

Subjects imagine what a piece of paper would look like when it is unfolded after having been folded and cut as shown. They then select one of the possible solutions, A, B, C, D, or E. From Thorndike, R. L., Hagen, & E. P., Sattler, J. M. ([1]986). *The Stanford-Binet scale of intelligence*. Riverside, IL: Chicago.

Figure 2: An Item from Raven's Progressive Matrices

Subjects select one of the possible solutions to complete the pattern shown in the three by three matrix. From Gregory, R. J. (1996). *Psychological testing: History, principles, and applications (2nd ed).* Boston: Allyn and Bacon.

Figure 3: An Item from the Pattern Analysis Subtest of the Stanford-Binet: Fourth Edition. Subjects have several cubes. The cubes have designs on their six sides as shown in (a). Subjects use their cubes to create images such as the one shown in (b). From Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *The Stanford-Binet xcale of intelligence*. Riverside, IL: Chicago.

Figure 4: An Item from the Object Assembly Subtest of the Wechsler Preschool and Primary Scales of intelligence, Revised.

Subjects assemble a simple puzzle without a model of the completed image. From Wechsler, D. (1967). *Manual for the Wechsler preschool and primary scale of intelligence*. New York: The Psychological Corporation.

Figure 5: Rising Effect Size *r*s for the relationship between SAT scores and 4 years of arts courses compared to no years, 1998-1999 (1993 missing).

Figure 1

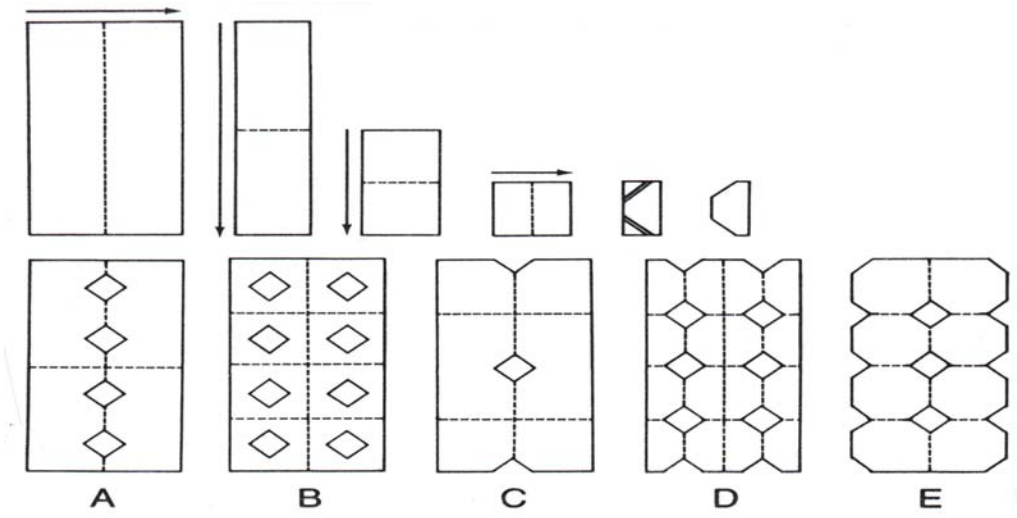

Figure 2

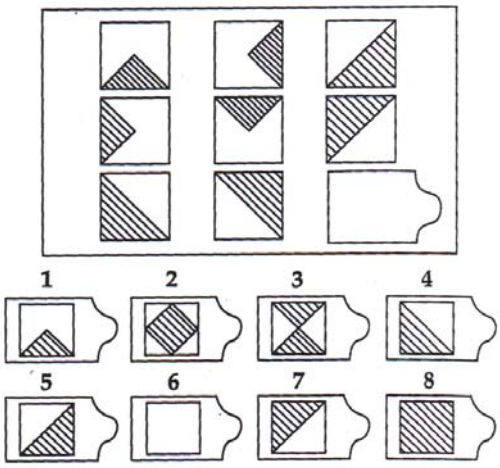

Figure 3a - b

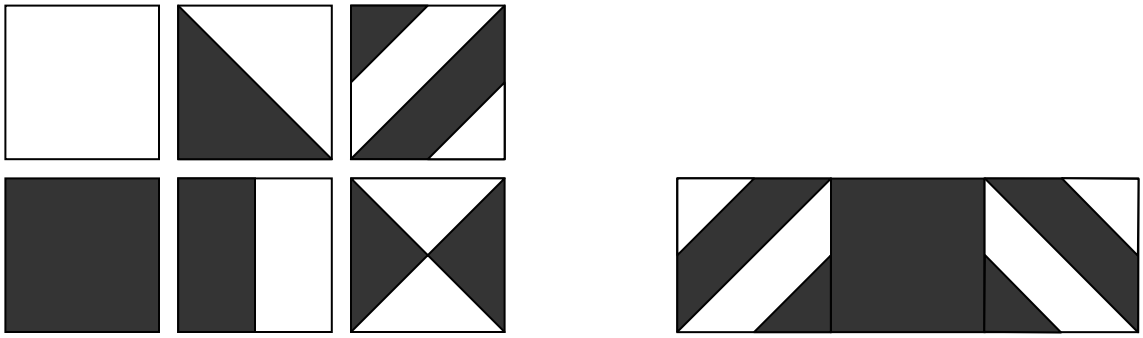

Figure 4