Research Report

THE MYSTERY OF THE MOZART EFFECT: Failure to Replicate

Kenneth M. Steele, Karen E. Bass, and Melissa D. Crook
Appalachian State University

Abstract—*The Mozart effect is the purported increase in spatial-reasoning performance immediately after exposure to a Mozart piano sonata. Several laboratories have been unable to confirm the existence of the effect despite two positive reports from the original laboratory. The authors of the original studies have provided a list of key procedural components to produce the effect. This experiment attempted to produce a Mozart effect by following those procedural instructions and replicating the procedure of one of the original positive reports. The experiment failed to produce either a statistically significant Mozart effect or an effect size suggesting practical significance. This general lack of effect is consistent with previous work by other investigators. We conclude that there is little evidence to support basing intellectual intervention programs on the existence of the Mozart effect.*

Rauscher, Shaw, and Ky (1993) reported that 36 undergraduates increased their mean spatial-reasoning scores the equivalent of 8 to 9 IQ points on portions of the Stanford-Binet Intelligence Scale: Fourth Edition (Thorndike, Hagen, & Sattler, 1986) after listening to 10 min of Mozart's Sonata for Two Pianos in D Major, K. 448. (Hereafter, we refer to this effect as the "Mozart effect.") The Mozart effect was temporary, having disappeared within 10 to 15 min. Subsequently, Rauscher, Shaw, and Ky (1995) reported a replication of the Mozart effect, using elaborations of the Stanford-Binet Paper Folding and Cutting (S-B PF&C) task as the dependent measure.

The hypothesis that musical experiences of short duration can have a direct causal influence on spatial reasoning on both a short-term and a long-term basis (Rauscher, 1997) is important for both practical and theoretical reasons. However, several attempts by other laboratories to confirm the existence of a Mozart effect have been unsuccessful (Carstens, Huskins, & Hounshell, 1995; Dalla Bella, Dunlop, Dawe, Humphrey, & Peretz, 1999; Kenealy & Monsef, 1994; Newman et al., 1995; Steele, Ball, & Runk, 1997; Stough, Kerkin, Bates, & Mangan, 1994; Weeks, 1996).

Rauscher and Shaw (1998) reviewed some of the negative results and described key components necessary to produce a Mozart effect. Negative results in some studies were explained by the choice of an inappropriate dependent measure, and the use of a PF&C task was endorsed. Rauscher and Shaw warned investigators to attend to issues concerning the order of presentation of the listening and task conditions. A major concern was that a pretest immediately before the treatment may produce a carryover effect that obscures enhancement by the music.

The purpose of this experiment was to confirm the existence of the Mozart effect by following the advice of Rauscher and Shaw (1998). The experiment was designed to be a faithful replication of the central conditions of the Rauscher et al. (1995) experiment. Rauscher et al. tested the effect of exposure to Mozart's Sonata for Two Pianos (K. 448) against silence and against highly repetitive music (Philip Glass's *Music With Changing Parts*) in a mixed-groups design. We chose to replicate the 1995 experiment because more procedural details were available for that experiment than for the earlier experiment, because the overall magnitude of the Mozart effect was stronger in the 1995 experiment, and because the procedure was consistent with the recommendations of Rauscher and Shaw (1998).

METHOD

Participants

Rauscher et al. (1995) employed 79 participants, distributed among three conditions. This study involved 125 participants (42 male and 83 female) distributed among three conditions. Participants came from introductory psychology courses and received credit for participation.

Materials and Apparatus

Rauscher et al. (1995) used several sets of 16 PF&C items that were derived from the Stanford-Binet IQ measure. (The actual test contains 2 practice items and 18 test items.) We used two sets of 16 PF&C items; each set was composed approximately of half derived items and half true items from the Stanford-Binet measure. The PF&C sets were of approximately equal difficulty and had been developed in another laboratory (Rideout & Laubach, 1996).[1] Each PF&C item was adjusted in size to occupy the center portion of an overhead transparency measuring 21.5 by 28 cm. Figure 1 shows an example PF&C item.

Stimulus tapes were created from the CD performances used in Rauscher et al. (1995). Mozart's Sonata for Two Pianos, K. 448 (Mozart, 1985, track 1) is a lively three-movement piece. Although invariably described as 10 min in duration, the first section *(Allegro con spirito)* is actually 8 min, 24 s in duration in this performance. Philip Glass's (1973, track 1) *Music With Changing Parts* is more than 60 min in duration without break; the first 8 min, 24 s was recorded for use in the experiment. The Glass composition eschews a traditional melody and uses repetition of units. The units are often only a few seconds long, and the repetition often lasts for several minutes.

Exposure to music is an established mood-induction technique, and musical selections by Mozart have been used to induce a mood of elation (Kenealy, 1997; Westerman, Spies, Stahl, & Hesse, 1996). It has been suggested that the performance difference interpreted as evidence of a Mozart effect could be produced indirectly through differences in

Address correspondence to Kenneth M. Steele, Department of Psychology, Appalachian State University, Boone, NC 28608; e-mail: steelekm@appstate.edu.

1. We thank Bruce Rideout for permission to use the sets.

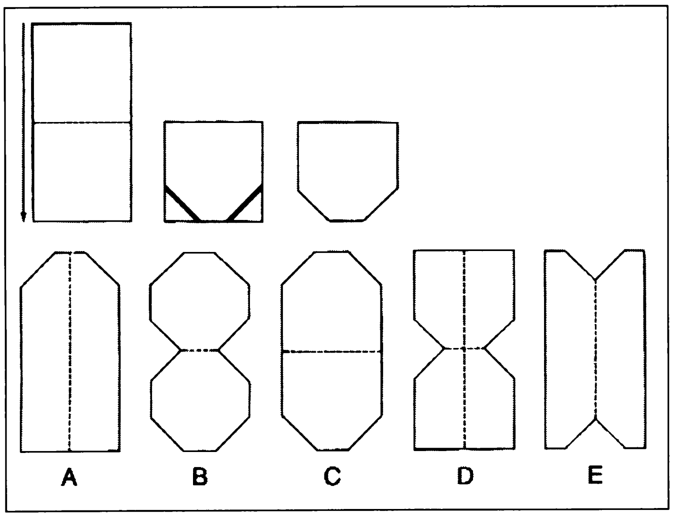

Fig. 1. Practice Stanford-Binet Paper Folding and Cutting item. The top row shows a piece of paper undergoing a series of transformations from left to right. The dotted line indicates the location of a fold; double lines indicate the location of cuts. The task is to pick the illustration in the bottom row that represents the appearance of the transformed paper when it is unfolded. For the item illustrated here, the correct answer is "C."

mood or arousal among treatments (Steele et al., 1997). We therefore used a mood measure to investigate this possibility. Individuals described the Mozart selection as "lively," "bouncy," and "happy"; the same individuals described the Glass selection as "repetitive," "obnoxious," and "grating." We hypothesized that such different reactions would be captured in a measure of mood. The 65-question Profile of Mood States (POMS; Educational & Industrial Testing Service, 1971) was adapted for use in this experiment by drawing 3 questions from each of the six mood factors reported by the measure. The factors are Depression, Tension, Anger, Vigor, Fatigue, and Confusion. Two questions unrelated to the POMS were added, 1 to begin and 1 to end our measure, for a total of 20 questions on mood, all answered using a 5-point ordinal scale. Our prediction was that the Glass selection would produce stronger indications of unpleasant mood relative to the Mozart sonata.

Tapes for the Mozart and the Glass selections were prepared from the CD recordings and were played on a Sony CFD-545 unit.

Procedure

Rauscher et al. (1995) used the following experimental procedure. First, all subjects were administered 16 PF&C items as a pretest. On the basis of their performance, they were assigned to create three groups of "equivalent capabilities." The next day, the three groups were exposed to 10 min of a stimulus condition and immediately tested with 16 PF&C items. The stimulus conditions consisted of listening to either the Mozart or the Glass selection or sitting in silence. Each PF&C item was shown via an overhead projector for 1 min, with a 5-s warning of the end of that trial. The three groups repeated this daily procedure for 3 additional days, with the exception that the group that heard the Glass selection heard other material on succeeding days.

Our procedure was slightly different. Because the Mozart effect was significant only on the 1st posttreatment day in the study by Rauscher et al. (1995), we restricted our experiment to the one posttreatment assessment. Performance on the PF&C pretest was not scored prior to group assignment. Instead, we assumed that random assignment would create equivalent groups. For schedule reasons, a time period of 48 hr elapsed between sessions for our participants. Rauscher and Shaw's (1998) concern was that the pretest should not occur too close in time to the treatment condition; consequently, our lengthening of the interval by an additional 24 hr between pretest and treatment condition should not have affected the results adversely.

Sessions were conducted in the early evening when the psychology building was quiet. Participants were recruited to create a group of 15 students per session. Participants arrived at the first session and were informed that they would be participating in a "puzzle experiment." The two sample PF&C items from the Stanford-Binet IQ measure were used to explain the task. The first sample item was projected onto a white screen measuring 3.5 m by 3.5 m. The experimenter explained the task using instructions slightly modified and abbreviated from the instructions provided by the Stanford-Binet measure. The second sample S-B PF&C item was then presented and explained. After the participants were given the opportunity to ask questions, answer sheets were provided, and 16 PF&C items were projected for up to 1 min each (depending on how quickly all subjects had made their choice), with a 5-s warning of the end of each 1-min period. The number of participants was limited to approximately 15 per session to ensure that distance from the screen and visibility of the projected PF&C item were equal for all participants.

The second session began 48 hr later. Participants were reminded of the nature of the task, and answer sheets were distributed. Following exposure to the scheduled stimulus condition, participants were immediately tested on a new set of 16 PF&C items. The two sets of PF&C items were used in counterbalanced order across sessions and groups. After completing the PF&C task, the participants were immediately given the mood assessment instrument and instructed to indicate their mood when the PF&C task began. Performance on the PF&C items and answers to the mood questions were analyzed at a later date.

RESULTS

Table 1 shows the results of the experiment. The results are grouped according to subjects' assignment to treatment condition, but pretest results show performance on PF&C items prior to the treatment condition. The pretest results indicate that random assignment was

Table 1. *Mean number of Paper Folding and Cutting items answered correctly*

		Pretest		Posttreatment	
Listening condition	*N*	*M*	*SE*	*M*	*SE*
Mozart	44	9.66	0.56	11.77	0.48
Silence	42	9.88	0.47	11.60	0.43
Glass	39	9.90	0.70	12.15	0.62

successful in creating groups not significantly different in initial task performance, $F(2, 122) = 0.05$, $p = .95$.

The posttreatment results, the mean number of PF&C items chosen correctly after exposure to the scheduled treatment condition, indicate that improvement in performance was approximately equal for all groups. This interpretation is supported by the results of analysis of variance: There was a significant main effect of session, $F(1, 122) = 76.1$, $p < .001$, but no significant effect of treatment, $F(2, 122) = 0.11$, $p = .89$, and no significant interaction of treatment by session, $F(2, 122) = 0.48$, $p = .62$.

A more sensitive measure is to assess relative change at the level of the individual. An analysis of covariance was used to examine post-treatment performance, adjusted for an individual's initial performance on the PF&C task. There was no significant treatment effect observed with this procedure either, $F(2, 121) = 0.61$, $p = .55$.

Treatment condition did not influence cognitive task scores but did influence mood scores in the anticipated manner. Significant differences among the groups were seen for both the Tension factor, $F(2, 122) = 6.32$, $p = .002$, and the Anger factor, $F(2, 122) = 7.21$, $p = .001$. Tension scores were highest for the Glass group, intermediate for the silence group, and lowest for the Mozart group (Tukey HSD pair-wise probabilities: Glass vs. Mozart, $p = .001$; Glass vs. silence, $p = .075$). Anger scores were also highest for the Glass group, intermediate for the silence group, and lowest for the Mozart group (Tukey HSD pair-wise probabilities: Glass vs. Mozart, $p = .001$; Glass vs. silence, $p = .019$).

DISCUSSION

The main result was that no significant Mozart effect was found despite replication of the procedure used by Rauscher et al. (1995). There is a large discrepancy between the results of the two studies. The effect size for the contrast of Mozart versus silence in Rauscher et al. (1995) was substantial ($d = 0.72$), whereas the effect size for the same contrast in this study was $d = 0.06$. One would have expected that fidelity to the procedure of Rauscher et al. would produce a similar effect size. Chabris (1998) calculated an average effect size of $d = 0.16$ for all 15 Mozart-versus-silence comparisons published, or submitted, to date.

Although there was no Mozart effect on cognitive task performance, there was an effect on mood scores. Participants did attend to the music, and were significantly less happy listening to the amelodic, repetitive selection from Glass than listening to silence or the Mozart sonata. Because differences in mood have been shown to affect performance on other cognitive tasks (Kenealy, 1997; Spies, Hesse, & Hummitzsch, 1996), these results suggest that production of a performance difference indirectly through differences in mood or arousal must be differentiated from the direct neurophysiological priming effect hypothesized in Rauscher et al. (1993).

If mood or arousal differences are important, then the presence of other subjects in a group testing procedure may have effects. In a replication of the present experiment with the modification that only one individual was tested in a session, the results were the same (Steele, Shannon, Kirby, & Olmstead, 1998). There was no differential improvement on the spatial-reasoning task, although treatments did affect mood, as in the present study.

There are issues beyond negative empirical results that confront the investigator of the Mozart effect. One issue concerns the lack of specification of the class of music selections that are supposed to have the hypothesized effect. Rauscher and Shaw (1998) have not advanced beyond descriptions such as "complexly structured music" (p. 839). Rauscher and Shaw cited Nantais (1997); Rideout, Dougherty, and Wernert (1998); and Wilson and Brown (1997) as providing confirmation of the Mozart effect. However, all three experiments found that music selections other than the Mozart sonata produced significant improvement relative to the control condition. For example, Rideout et al. found improvement after subjects listened to a performance by Yanni. It is difficult to determine whether this musical selection is or is not consistent with the "complex structure" criterion of Rauscher and Shaw. Additionally, Nantais found that the choice of control condition was a critical issue. Nantais reported a significant difference when the control condition was sitting alone in silence, but not when the control condition was listening to a story. This pattern of results is consistent with an account focused on mood or arousal differences among conditions.

Another issue is uncertainty regarding the suitability of dependent measures. In the original report, Rauscher et al. (1993) presented only the combined performance from the three Stanford-Binet subtests (PF&C, Matrices, and Pattern Analysis). In that report, they stated, "For our sample, these three tasks correlated at the .01 level of significance. We were thus able to treat them as equal measures of abstract reasoning ability" (p. 611).

The results of the 1993 experiment were recounted differently in Rauscher and Shaw (1998). Rauscher and Shaw conjectured that other investigators have had difficulties in producing a Mozart effect because the investigators have not distinguished between spatial-temporal tasks (like the PF&C task) and spatial pattern-recognition tasks (like the Raven Progressive Matrices task). Rauscher and Shaw indicated that their 1993 effect occurred only with the S-B PF&C task, not with the Matrices or Pattern Analysis tasks. This improvement difference among the three Stanford-Binet tasks was used to account for the negative results obtained from studies that used the Raven Progressive Matrices task. However, no such difference in Stanford-Binet task performance was found by Dalla Bella et al. (1999) and Kenealy and Monsef (1994), who used both the S-B PF&C and the Matrices tasks, or by Weeks (1996), who used all three tasks reported in Rauscher et al. (1993).

The popular excitement about the Mozart effect is due to its claim to be a brief, easy remedy to improve intellectual skills. However, previous attempts to increase IQ scores demonstrate how difficult it is to produce even a small, short-lived gain (Spitz, 1986). Considering the duration and depth of many intervention programs, an effect from short periods of listening to music does not seem feasible in principle. The results of this experiment, and experiments in other laboratories, are consistent with such an expectation. We conclude that there is little evidence to support basing intellectual enhancement programs on the existence of the causal relationship termed the Mozart effect.

Acknowledgments—The work reported here was supported by grants from the Cratis D. Williams Graduate School, Appalachian State University. Portions of these results were submitted by Karen Bass as a thesis in partial fulfillment of the requirement for graduation with honors in psychology. Portions of these results were presented previously at meetings of the Southeastern Psychological Association, the North Carolina Cognition Conference, and the Psychonomic Society. We thank Stan Aeschleman for comments on the manuscript.

REFERENCES

Carstens, C.B., Huskins, E., & Hounshell, G.W. (1995). Listening to Mozart may not enhance performance on the revised Minnesota paper form board test. *Psychological Reports, 77,* 111–114.

Chabris, C.F. (1998). *Brief exposure to music does not increase intelligence.* Manuscript submitted for publication.

Dalla Bella, S., Dunlop, T., Dawe, L., Humphrey, K., & Peretz, I. (1999, April). *The Mozart effect revisited.* Paper presented at the annual meeting of the Cognitive Neuroscience Society, Washington, DC.

Educational & Industrial Testing Service. (1971). *Profile of Mood States.* San Diego: Author.

Glass, P. (1973). *Music with changing parts* [CD]. New York: Elektra Nonesuch.

Kenealy, P. (1997). Mood-state-dependent retrieval: The effects of induced mood on memory reconsidered. *Quarterly Journal of Experimental Psychology, 50A,* 290–317.

Kenealy, P., & Monsef, A. (1994). Music and IQ tests. *The Psychologist, 7,* 346.

Mozart, W.A. (1985). Sonata for two pianos, K. 448 [Recorded by M. Perahia & R. Lupu]. On *Mozart: Sonata, K.448; Schubert: Fantasia, op. 103, D.940* [CD]. New York: CBS.

Nantais, K.M. (1997). *Spatial-temporal skills and exposure to music: Is there an effect, and if so, why?* Unpublished master's thesis, University of Windsor, Windsor, Canada.

Newman, J., Rosenbach, J.H., Burns, K.L., Latimer, B.C., Matocha, H.R., & Vogt, E.E. (1995). An experimental test of "the Mozart effect": Does listening to his music improve spatial ability? *Perceptual and Motor Skills, 81,* 1379–1387.

Rauscher, F.H. (1997). A cognitive basis for the facilitation of spatial-temporal cognition through music instruction. In V. Brummett (Ed.), *Ithaca Conference '96: Music as intelligence: A sourcebook* (pp. 31–44). Ithaca, NY: Ithaca College.

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,* 611.

Rauscher, F.H., Shaw, G.L., & Ky, K.N. (1995). 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.

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

Spies, K., Hesse, F., & Hummitzsch, C. (1996). Mood and capacity in Baddeley's model of human memory. *Zeitschrift fur Psychologie, 204,* 367–381.

Spitz, H.H. (1986). *The raising of intelligence: A selected history of attempts to raise retarded intelligence.* Hillsdale, NJ: Erlbaum.

Steele, K.M., Ball, T.N., & Runk, R. (1997). Listening to Mozart does not enhance backwards digit span performance. *Perceptual and Motor Skills, 84,* 1179–1184.

Steele, K.M., Shannon, R.A., Kirby, J.L., Jr., & Olmstead, C.G. (1998). *Failure to produce a Mozart effect with individual testing.* Unpublished manuscript, Appalachian State University, Boone, NC.

Stough, C., Kerkin, B., Bates, T., & Mangan, G. (1994). Music and spatial IQ. *Personality and Individual Differences, 17,* 695.

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

Weeks, S.P. (1996). The effect of music on abstract/visual reasoning performance in high school music and non-music students. *Dissertation Abstracts International, 56*(9), 3408A. (UMI Dissertation Services No. 9600110)

Westerman, R., Spies, K., Stahl, G., & Hesse, F.W. (1996). Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology, 26,* 557–580.

Wilson, T.L., & Brown, T.L. (1997). Reexamination of the effect of Mozart's music on spatial-task performance. *The Journal of Psychology, 131,* 365–370.

(Received 6/28/98; Revision accepted 1/13/99)