WEAK JUSTIFICATION FOR COMBINING GROUPS
A Reply to Schellenberg (2005)

Kenneth M. Steele
Appalachian State University

Schellenberg (2004) assigned 144 6-year-old children to one of four treatments in a pretest-posttest design. Subjects experienced either keyboard, Kodály voice, drama, or no lessons over an academic year. Schellenberg reported a significant difference of 2.7 IQ points if the keyboard and voice results were combined and contrasted against the combined drama and no-lessons results. Steele (2005) and Black (2005) questioned the justification for this combination and reported that the general pattern was weak and nonsignificant differences in the absence of this particular combination. Schellenberg (2005) attempted to justify this combination by appeals to the expense of the project and public interest in the research question. The result is explained as being due to the school-like nature of the activities. This explanation raises the question of why drama lessons are not school-like and instead were combined with the no-lessons group.

Schellenberg (2004) assigned 144 6-year-old subjects to one of four treatment conditions in a pretest-posttest design. Subjects experienced either piano-keyboard lessons, voice lessons accompanied by clapping and rhythmic movement, drama lessons, or no lessons. The primary measure was IQ gain scores over the course of a year. Schellenberg reported a small, significant difference of 2.7 IQ points if the keyboard and voice groups were combined and contrasted against the combined results of the drama and no-lessons groups. Steele (2005) questioned the justification for combining the keyboard and the voice groups; and Black (2005) questioned the justification for combining the drama and no-lesson groups. Steele and Black found that statistical significance disappeared in the absence of this combination. Both Steele and Black reported that the general pattern is that of weak and insignificant effects unless one uses the particular combination of groups reported in Schellenberg (2004). The central question is whether reporting *only* the statistical results from the combined groups misrepresents the general patterns of results. Here I respond to Schellenberg's comments.

Correspondence concerning this article should be addressed to Kenneth M. Steele, Department of Psychology, Appalachian State University, Boone, NC 28608; e-mail: steelekm@appstate.edu.

After looking at the results, Schellenberg (2005) declares the combinations of keyboard plus voice lessons and drama plus no-lessons to be obvious. This is the beauty of post hoc reasoning: effects occur as postdicted. However, if one had to make predictions about appropriate post hoc combinations in advance of the results, there are competing alternatives. For example, Kodály and drama lessons involve the use of expressive movement. Perhaps they should have been combined into an "expressive movement" group to be compared against keyboard lessons. Or perhaps keyboard, voice, and drama should have been combined into an "arts" group to be contrasted against the no-lessons group. There are many possible combinations. Which ones are justified? Schellenberg provides no strong theoretical justification for his combination other than "the magnitude and expense of the project, combined with considerable public and scientific interest in the research question, meant that it was essential to disseminate the findings as widely as possible."

My general approach was to ask how one should think about the four treatments. Analyses should be based on the hypothesized effects of the treatments rather the cost of the study. My first approach was that the four groups were exposed to different and independent experiences. Therefore, one standard technique would have been to do a one-way ANOVA on the gain scores across the treatments, followed by pair-wise contrasts to determine which treatments were significantly different. The results of this analysis were not significant, the overall effect was weak ($\eta^2 = .03$), and *none* of the pair-wise contrasts approached statistical significance.

Schellenberg's reply is that he didn't report the analysis because the result "was *not* significant, which should be obvious to the average reader of *Psychological Science* from the statistics that were reported." What is "obvious" to the average reader of *Psychological Science* is an empirical question. However, in my case, when I see an article titled "Music Lessons Enhance IQ," then I presume effects were clearly present. The fact that no single group was significantly different from any other single group is important information. Not being informed that statistical significance depends upon a specific post hoc combination of groups *does* misrepresent the general pattern of results.

My second approach was to treat the four groups as being ordered on a continuum of exposure to music lessons in order to use a regression analysis. I suggested that the groups be ordered keyboard, voice, drama, and no-lessons (most to least) in terms of exposure to music lessons. Schellenberg wants to know what made me "privilege" the keyboard group over the voice lessons group. Here I was following the empirical literature. Piano and keyboard instruction is the most common treatment in this type of study (Costa-Giomi, 1999, 2004; Rauscher, Shaw, Levine, et al. (1997); Rauscher & Zupan, 2000). Each report by Costa-Giomi (1999, 2004) involved a 3-year longitudinal study. Clearly researchers in this area think that there is something special about standard keyboard instruction.

The results of the regression analysis approached statistical significance ($p = .07$), but the effect was very weak ($R^2 = .03$). Schellenberg made mistakes in his comments about the regression analysis. First, he chastised me for assuming an equidistant interval scale for the music instruction dose. But I didn't make that assumption, and a linear regression analysis is not based on that assumption. Second, R^2 is a measure of accounted-for variance and not a Pearson correlation.

Schellenberg suggested the computation of a Spearman correlation as the "appropriate test for rank-ordered data." It is difficult to interpret these results when one considers the computational steps that occurred. A Spearman correlation shows the relationship between two sets of ranked data, so both IQ gain scores and music dose values will be converted into ranks. However there are only four values (1, 2, 3, 4) in the 132 cases of the music dose variable. Ranking of the music dose variable will produce massive ties resulting in a variable that contains 132 observations comprised of only four values. A polyserial correlation would have been the appropriate measure to use with an interval and ordinal variable.

Schellenberg's reply to Black's (2005) commentary is misleading in a very important manner. Schellenberg uses the plural repeatedly, as in "The music groups had increases in FSIQ exceeding those of the no-lessons group" The use of the plural suggests that both groups were significantly different from the no-lessons group. This is not the case. As I pointed out earlier, *no* pair-wise contrast among the four treatments approached statistical significance. Statistical significance can be obtained *only* if you combine the keyboard and the voice results.

Schellenberg's current interpretation of the results is that the effect is due to the school-like nature of the activities. This is a retreat from Schellenberg's (2004) conclusion that the result represents a rare case of *far* transfer effects of music lessons to IQ scores. Now he is faced with the problem of explaining why drama lessons should not be combined with voice and keyboard lessons into a "school-like" group. Why are drama lessons are not school-like enough? Schellenberg (this issue) informs us that drama lessons involve playlike activities such as *pretending* and *dressing up*. But Schellenberg stated earlier that he would have predicted voice lessons to have a greater impact than standard conservatory keyboard lessons because they were more child-centered and enjoyable. It does not seem plausible that play-like enjoyment both increases IQ gains of voice lessons over keyboard lessons and reduces IQ gains in the drama group. The combined results of the keyboard, voice and drama groups was not significantly different from the no-lessons group, $t(130) = 1.42, p = .16$.

Steele (2005) made the point that "music lessons" may be a poor choice for an independent variable because the crucial experience itself needs description. Schellenberg's reply is that many constructs, like intelligence, are messy but worthy of investigation. I agree that investigation of what is meant by "music lessons" and its impact on children is clearly legitimate. But those questions were not the reason for this particular line of research. This study follows the path of research begun by Rauscher, Shaw, and Ky (1993) that claims a special causal relationship between music and certain types of intelligence (Rauscher, 2002). I suggest it is time to move on from these *Mozart effect* variations that use IQ benefits as an argument for music lessons. Music should be included in school curricula because of its intrinsic merits.

References

Black, S. L. (2005). Hath music charms to boost the average IQ? *The Scientific Review of Mental Health Practice*, *4*(2), 3–5.

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

Costa-Giomi, E. (2004). Effects of three years of piano instruction on children's academic achievement, school performance, and self-esteem. *Psychology of Music*, *32*, 139–152.

Rauscher, F. H. (2002). Mozart and the mind: Factual and fictional effects of musical enrichment. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 267–278). San Diego, CA: Academic Press.

Rauscher, F. H., Shaw, G. L., & Ky, K. N. (1993). Music and spatial task performance. *Nature*, *365*, 611.

Rauscher, F. H., Shaw, G. L., Levine, L. J., Wright, E. L., Dennis, W. R., & Newcomb, R. L. (1997). Music training causes long-term enhancement of preschool children's spatial-temporal reasoning. *Neurological Research*, *19*, 2–8.

Rauscher, F. H., & Zupan, M. A. (2000). Classroom keyboard instruction improves kindergarten children's spatial-temporal performance: A field experiment. *Early Childhood Research Quarterly*, *15*, 215–28.

Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychological Science*, *15*, 511–514.

Schellenberg, E. G. (2005). Music lessons enhance IQ: A reply to Black (2005) and Steele (2005). *The Scientific Review of Mental Health Practice*, *4*(2), 10–13.

Steele, K. M. (2005. Do Music Lessons enhance IQ? A reanalysis of Schellenberg (2004). *The Scientific Review of Mental Health Practice*, *4*(2), 6–9.