Standardization of the Gordon Primary Measures of Music Audiation in Greece

Lelouda Stamou, Charles P. Schmidt and Jere T. Humphreys
Journal of Research in Music Education 2010 58: 75
DOI: 10.1177/0022429409360574

MENC The National Association for MUSIC EDUCATION

Journal of Research in Music Education
58(1) 75–89

DOI: 10.1177/0022429409360574
http://jrme.sagepub.com

Standardization of the Gordon Primary Measures of Music Audiation in Greece

Lelouda Stamou[1], Charles P. Schmidt[2], and Jere T. Humphreys[3]

Abstract

The purpose of this study was to standardize the Primary Measures of Music Audiation in Greece (N = 1,188). Split-halves reliability was acceptable across grade levels (K through 3) for the Tonal and Rhythm subtests, but test–retest reliability was generally unacceptable, especially for the Rhythm subtest. Concurrent validity was mixed, with teacher ratings of musical achievement generally significantly correlated with Tonal but not Rhythm subtest scores. Composite test means were significantly higher for suburban and urban samples than for rural samples and were significantly higher for higher grade levels. Item difficulty coefficients were significantly correlated across grade levels. The Greek and U.S. composite means were similar except for a significantly higher U.S. mean for grade 1. However, when the rural subgroup was removed from the Greek sample to equate with the U.S. norming sample, there were nonsignificant differences for grades K through 1, but significant differences in favor of the Greek sample for grades 3 and 4.

Keywords

music aptitude tests, audiation, standardization, Greece, cross-cultural

The measurement of music aptitude has been of interest to researchers since the beginning of the 20th century. Emanating from the psychological research of Francis Galton (1822–1911) and James Cattell (1860–1944), interest in quantifying individual differences and devising discrete measures of sensory acuity led to Carl Seashore's music

[1]University of Macedonia, Thessaloniki, Greece
[2]Indiana University, Bloomington, IN, USA
[3]Arizona State University, Tempe, AZ, USA

Corresponding Author:
Jere T. Humphreys, School of Music, Arizona State University, Tempe, AZ 85287-0405
Email: Jere.Humphreys@asu.edu

aptitude tests, published in 1919 (Humphreys, 1993, 1998). In the second half of the 20th century, based on the work of Seashore (1919), his followers, and others (Bentley, 1966; Drake, 1933, 1954; Gaston, 1957; Kwalwasser, 1953, 1955; Kwalwasser & Dykema, 1930; Tilson, 1941; Whistler & Thorpe, 1950; Wing, 1958), Edwin Gordon conducted research that led to the development of four music aptitude tests: Primary Measures of Music Audiation (PMMA) for grades K through 3 (Gordon, 1979), Intermediate Measures of Music Audiation (IMMA) for children in grades 1 through 4 who score above the 80th percentile on the PMMA (Gordon, 1982), Musical Aptitude Profile (MAP) for grades 4 through 12 (Gordon, 1965), and Advanced Measures of Music Audiation (AMMA) for college-age students (Gordon, 1989).

These tests have been used in numerous research studies not only in the United States, where all four tests were normed, but also in other countries. However, LeBlanc, Jin, Stamou, and McCrary (1999) concluded that "research findings based upon one culture cannot safely be assumed to apply to other cultures" (p. 76). More specifically, Feay-Shaw (2000) questioned the lack of concern shown for differences in culture, ethnicity, and socioeconomic status in several research studies involving tests of music aptitude, especially in how participant groups were described and how results were generalized and otherwise interpreted. Later, on the basis of results of a study of the effects of beat strength on music listening preference, LeBlanc et al. (2002) concluded that a variable called "country" or "culture" should be added to the LeBlanc (1980) preference model.

Several researchers have investigated how Gordon's music aptitude tests functioned in various countries or cultures. Chung (2002) reported comparable reliability coefficients for the IMMA with Korean students in Korea and Korean American students (grades 1 through 4). Schoenoff (1972) found no significant differences between MAP scores of German participants and the published norms (grades 4 through 12), but Sell (1976) and Chuang (1997) reported significantly higher scores than the MAP norms for Finnish and Taiwanese students, respectively, as well as high reliability and concurrent validity coefficients. Jung (1992) conducted a 2-year predictive validity study of the MAP for use in Korea and found that MAP scores correlated significantly with a measure of Western music achievement ($r = .33$ to $r = .66$). In contrast, MAP scores had low or nonsignificant correlations with a measure of achievement in Korean traditional music ($r = .01$ to $r = .28$). Özeke and Humphreys (2000) and Wang (2007) reported significantly higher scores on the AMMA than the published norms for Turkish and Taiwanese students, respectively. Gordon (1991) validated the AMMA specifically with and for German students.

In foreign applications of the PMMA, Yang (2002) reported that first-grade students in Taiwan scored significantly lower than the published norms on the Tonal subtest and nonsignificantly lower on the Rhythm subtest. Yang reported split-halves reliability coefficients ranging between .84 and .91. Examining the PMMA with a sample of English children, Holahan and Thomson (1981) found adequate split-halves reliability ($r = .86$ to $r = .93$) for the Tonal subtest by age group. In contrast, split-halves reliability coefficients for the Rhythm subtest ranged from unacceptable ($r = .45$) to marginally acceptable ($r = .71$). Gouzouasis (1993) found significant differences in

PMMA Tonal subtest scores, but not in Rhythm subtest scores, among groups of 5-year-old Canadian children of Chinese, East Indian, and Western European ethnic backgrounds. Gouzouasis reported high split-halves reliability ($r > .90$) for both Tonal and Rhythm subtests for each subgroup. However, the author concluded that content validity for the Tonal subtest "may be called into question for use with East Indian children" (p. 75), who scored lower than those in the other groups. The current investigation focused on use of the PMMA in Greece. One study (Pollatou, Karidimou, & Gerodimos, 2005) involved use of the PMMA in Greece with a sample of 5- and 6-year olds ($N = 95$). The authors reported descriptive results for the PMMA as follows: Tonal subtest ($M = 22.68$, $SD = 5.78$), Rhythm subtest ($M = 22.62$, $SD = 4.64$). No reliability or validity data were reported.

When standardized music tests are applied to populations other than the ones for which the tests were constructed and normed, the crucial issue of validity emerges. Because the validity and reliability of measurement instruments cannot be assumed to transfer across populations, empirical data are needed to determine the viability of measurement instruments, particularly when used cross-culturally. Therefore, the purpose of this study was to standardize a well-known music aptitude test for use in Greece: the PMMA by Edwin E. Gordon (1979). The two main research objectives were to (a) develop PMMA norms appropriate for use in Greece and (b) determine the reliability and concurrent validity of each subtest at each grade level. Other research objectives were to (a) compare results by grade (K through 3) and type of school setting (i.e., urban, suburban, rural) and (b) compare results from Greece with those from Gordon's original norming study.

Method

The written instructions for the PMMA Tonal and Rhythm subtests were translated from the original English into Greek by the primary author of this study, a native Greek with fluency in English. She also replaced the English-language instructions on the audio test CDs with Greek translations.

The main standardization sample for this study included nearly equal numbers of students from 12 urban schools ($n = 397$, 33.4%), 14 suburban schools ($n = 391$, 32.9%), and 16 rural schools ($n = 399$, 33.6%), all from Greece ($N = 1{,}188$). Participating schools were classified as urban, suburban, or rural according to an official demographic map published by the Greek government (General Secretariat of National Statistical Service of Greece, 2008). This was a nonrandom sample determined by accessibility (i.e., teachers' interest and willingness to participate) to the three types of schools and adequate class sizes. Every attempt was made to obtain a geographically diverse (including Greek Islands) and demographically varied sample. Class size ranged from 10 to 26 students. The male:female ratios were also nearly equal for grade 1 ($n = 86{:}100$), grade 2 ($n = 173{:}161$), and grade 3 ($n = 293{:}271$); however, the kindergarten sample consisted of 38 males (36.9%) and 65 females (63.1%). The overall usable sample (grades K through 3) consisted of 591 males (49.75%) and 597 females (50.25%). The

aptitude test was administered by the regular music teachers, all of whom had been trained in music aptitude test administration as part of a 20-clock-hour seminar in quantitative research and music aptitude testing organized and taught by the three present authors (Stamou, Humphreys, & Schmidt, 2006).

As a parallel to Gordon's (1979) methodology, we employed split-halves and retest reliability procedures. Samples employed for the retest reliability calculations for each grade and subtest or composite ranged from $n = 31$ for the grade 1 composite (i.e., Tonal and Rhythm subtests combined) to $n = 112$ for the grade 2 Rhythm subtest. The time interval between test administrations ranged from 10 days to 4 weeks.

For purposes of assessing concurrent criterion validity, we asked music teachers to provide ratings of individual students' music aptitude and music achievement based on their observations of individual students across time. The researchers provided teachers with written and verbal explanations concerning the definitions of music aptitude (i.e., perceived potential for future success in music) and music achievement (i.e., current musical competence or behavior) and were asked to rate each student on a 7-point scale (1 = *low,* 7 = *high*) for each category. The practice of correlating teacher ratings with test scores is similar to methodology employed by Gordon (1965) in his validation of the MAP. However, this differed from Gordon's (1979) validation procedures for the PMMA, as he assessed validity by correlating scores with standardized measures of academic achievement for each grade level and by correlating PMMA and MAP scores for a sample of fourth graders.

Results

Descriptive statistics and reliability coefficients for each subtest and composite are presented by grade level in Table 1. Split-halves reliability coefficients for the Tonal subtest ranged from moderate for the kindergarten sample ($r = .77$) to moderately high for grades 1, 2, and 3 ($r = .86$ to $.88$). Split-halves coefficients for the Rhythm subtest ranged from unacceptable levels for kindergarten ($r = .42$) and grade 1 ($r = .50$) to marginally acceptable levels for grade 2 ($r = .66$) and grade 3 ($r = .69$). Split-halves reliability coefficients for the composite measure were acceptable, ranging from $r = .73$ for kindergarten to $r = .86$ for grade 3 (Nunnally, 1970). However, the relatively low reliability of the Rhythm subtest accounts for the unusual finding that for each grade level, the 40-item Tonal subtest was more reliable than the 80-item composite (Table 1).

Retest reliability results for the PMMA Tonal, Rhythm, and composite measures varied markedly by grade level and by subtest but were generally unacceptable, with Rhythm subtest scores being particularly unstable across time (Table 1). We excluded the kindergarten retest data because of inadequate sample sizes (retest $n = 11$ to 15). The highest retest reliability was found for grade 1 Tonal ($r = .54$), Rhythm ($r = .45$), and composite ($r = .64$). There was a modest degree of stability in scores for grade 2 Tonal ($r = .42$) and composite ($r = .50$) but very low stability for grade 2 Rhythm ($r = .19$). Grade 3 reliability results were even lower, ranging from $r = -.11$ for Rhythm to $r = .10$ for the composite test.

Table 1. Primary Measures of Music Audiation Means, Standard Deviations, and Reliabilities

	Current study			Gordon (1979)		
		Reliability			Reliability	
Grade and measure	*M* (*SD*)	Split halves	Retest	*M* (*SD*)	Split halves	Retest
Kindergarten						
Tonal (*n* = 126)	22.95 (5.33)	.77	—	24.70 (5.28)	.88	.73
Rhythm (*n* = 129)	23.46 (4.22)	.42	—	22.30 (3.74)	.72	.60
Composite (*n* = 103)	46.56 (7.94)	.73	—	47.00 (7.65)	.90	.74
Grade 1						
Tonal (*n* = 224)	27.68 (5.50)	.86	.54 (*n* = 52)	29.80 (5.03)	.89	.70
Rhythm (*n* = 222)	25.94 (4.41)	.50	.45 (*n* = 37)	25.80 (4.34)	.85	.66
Composite (*n* = 186)	53.63 (8.81)	.84	.64 (*n* = 31)	55.60 (8.25)	.92	.75
Grade 2						
Tonal (*n* = 358)	31.22 (5.41)	.80	.42 (*n* = 99)	32.00 (4.75)	.89	.70
Rhythm (*n* = 398)	29.23 (4.76)	.66	.19 (*n* = 112)	27.70 (4.55)	.86	.73
Composite (*n* = 334)	60.87 (8.50)	.77	.50 (*n* = 83)	59.70 (8.35)	.92	.76
Grade 3						
Tonal (*n* = 620)	32.90 (5.04)	.88	–.01 (*n* = 64)	34.60 (3.35)	.85	.68
Rhythm (*n* = 652)	31.05 (4.58)	.69	–.11 (*n* = 104)	29.40 (3.99)	.86	.66
Composite (*n* = 564)	64.05 (8.45)	.86	.10 (*n* = 50)	64.00 (6.29)	.90	.73

Split-halves reliability coefficients were calculated on even and odd items with Spearman-Brown correction. Retest reliability coefficients for kindergarten are not shown because of small (retest) sample sizes.

As a measure of concurrent validity, teachers' ratings of individual students' musical achievement were correlated with student scores on the PMMA subtests and composite for grades 1 through 3 (Table 2). There were significant correlations for the Tonal subtest for grades 1 and 3 and for the composite test for grade 1 (*rho* = .39 to .55). By contrast, teachers' achievement ratings and Rhythm subtest scores were not significantly correlated at any grade level (*rho* = .08 to .21). An additional measure of concurrent validity, teachers' ratings of students' musical aptitude, was not significantly correlated with subtest or composite test scores at any grade level (Table 2).

Intercorrelations between the Tonal and Rhythm subtests ranged from $r = .43$ for kindergarten to $r = .62$ for grades 1 and 3 (Table 3). Correlations between the Tonal and composite scores ranged from $r = .89$ to $r = .92$ and between Rhythm and composite scores from $r = .80$ for kindergarten to $r = .89$ for grades 2 and 3. All correlations were statistically significant ($p < .001$) because of the large sample sizes involved.

Given the high correlations between the PMMA subtests and composite, only the composite scores were compared for school settings and grade levels. The assumption

Table 2. Spearman Correlations (*rho*) for Teachers' Musical Achievement and Aptitude Ratings With PMMA Tonal, Rhythm, and Composite Scores

Grade and measure	*n*	Musical achievement rating	Musical aptitude rating
Grade 1			
Tonal	21	.53*	.23
Rhythm	21	.21	–.08
Composite	21	.55*	.24
Grade 2			
Tonal	21	.33	.23
Rhythm	21	.18	.02
Composite	21	.21	.03
Grade 3			
Tonal	74	.39***	.17
Rhythm	65	.08	.01
Composite	65	.25	.11

PMMA = Primary Measures of Music Audiation (Gordon, 1979).
*$p < .05$. ***$p < .001$.

Table 3. PMMA Intercorrelations (*r*) Among Subtests and Composite

	Current study		Gordon (1979)	
Subtest	Rhythm	Composite	Rhythm	Composite
	Kindergarten ($n = 103$)		Kindergarten ($n = 127$)	
Tonal	.43***	.89***	.45***	.76***
Rhythm	—	.80***	—	.57***
	Grade 1 ($n = 186$)		Grade 1 ($n = 202$)	
Tonal	.62***	.92***	.49***	.84***
Rhythm	—	.87***	—	.82***
	Grade 2 ($n = 334$)		Grade 2 ($n = 280$)	
Tonal	.62***	.91***	.51***	.89***
Rhythm	—	.89***	—	.88***
	Grade 3 ($n = 564$)		Grade 3 ($n = 264$)	
Tonal	.59***	.90***	.49***	.85***
Rhythm	—	.89***	—	.89***

PMMA = Primary Measures of Music Audiation (Gordon, 1979).
***$p < .001$.

of homogeneity of variance was met for the composite test among the urban, suburban, and rural school settings (Levene's $F = 1.74$, $p > .05$). An ANOVA test revealed significant overall differences among school settings, $F(2, 1187) = 9.78$, $p < .001$, but the effect size was small ($\eta^2 = .016$). Post hoc (Bonferroni) tests revealed significantly higher means for suburban ($p < .004$) and urban ($p < .000$) schools than for rural schools.

ANOVA results confirmed that the composite means were significantly related to grade level, $F(3, 1186) = 165.24$, $p < .001$, $\eta^2 = .30$. Because of lack of homogeneity of variance between grade levels (Levene's $F = 3.63$, $p < .05$), the Dunnett T3 multiple range test was used for grade-level comparisons. These results confirmed significant differences in composite means between all pairs of grade levels, with means increasingly higher for each grade level ($p < .001$).

We determined item difficulty and discrimination values for the Tonal and Rhythm subtests by grade level. For both subtests, as would be expected, mean difficulty coefficients were higher for the Tonal ($M = .57, .69, .78, .82$) and Rhythm ($M = .59, .64, .73, .78$) subtests for each grade, kindergarten and grades 1, 2, and 3, respectively. Similarly, the ranges of the difficulty coefficient means generally were larger for the Tonal (.65, .69, .78, .71) and Rhythm (.62, .72, .76, .81) subtests for grades K through 3, respectively.

The mean item discrimination values for the Tonal subtest suggest modest, decreasing levels of discriminatory power between high and low scorers for grades K through 3, respectively ($r = .36, .33, .30, .24$), with similarly decreasing ranges in the discrimination coefficients (.82, .76, .60, .43). The mean discrimination values for the Rhythm subtest were considerably higher and positively associated with grade level ($r = .59, .64, .73, .78$) for grades K through 3, respectively, with similar ranges among grade levels (.65, .65, .65, .52). The Rhythm subtest items tended to differentiate well among overall high and low scorers.

To address the degree to which individual items were consistent in their difficulty across grade levels, we calculated Spearman correlations among item difficulty values for the Tonal and Rhythm subtests by grade level. The difficulty of individual Tonal subtest items tended to be consistent ($p < .001$) across grade levels, with correlations (*rho*) ranging from .81, between kindergarten and grade 2 and between grades 1 and 2, to .93, between grades 2 and 3. The range of correlations for item difficulty values for the Rhythm subtest was somewhat wider: .59 between kindergarten and grade 3 and .92 between kindergarten and grade 1. Overall, there was a tendency toward relatively comparable degrees of difficulty across grade levels for individual Tonal and Rhythm items. That is, items that were relatively difficult to discriminate at one grade level were also relatively difficult to discriminate at another grade level.

Comparison With American Norms

Norms for the PMMA developed in this study (see Appendix) were based on a large, cross-sectional sample of schools, teachers, and students from urban, suburban, and rural school settings in Greece, whereas the original norms were based on a somewhat

smaller group of participants from a single American suburb more than 30 years ago (Gordon, 1979). The sample sizes by grade level in the current study approximate or exceed those from the original standardization study (kindergarten, $n = 103$ vs. 127; grade 1, $n = 186$ vs. 202; grade 2, $n = 334$ vs. 280; grade 3, $n = 564$ vs. 264; total, $n = 1{,}188$ vs. 873, respectively). Gordon did not provide data on the gender of the norming sample participants, although he did indicate that the sample was demographically heterogeneous.

The PMMA composite means for the U.S. (Gordon, 1979) and Greek samples were not significantly different for kindergarten ($t = .42$, $df = 228$, $p > .05$, $\eta^2 = .06$), grade 2 ($t = -1.71$, $df = 612$, $p > .05$, $\eta^2 = .14$), or grade 3 ($t = -.08$, $df = 826$, $p > .05$, $\eta^2 = .01$).[1] The U.S. composite mean was significantly higher for grade 1 ($t = 2.27$, $df = 386$, $p < .01$, $\eta^2 = .24$). Because the U.S. sample was drawn only from a suburban setting and because the Greek samples differed significantly between urban and suburban samples on one hand and the rural sample on the other, the authors then compared the U.S. composite means (Gordon, 1979) with the respective grade-level means from the combined urban and suburban Greek samples only, leaving out the Greek rural sample. For kindergarten, the mean for the truncated Greek sample was higher than the mean for the total Greek sample but remained nonsignificantly lower than the U.S. sample mean ($t = .06$, $df = 210$, $p > .05$, $\eta^2 = .01$). For grade 1, the significant difference in favor of the U.S. sample was reduced to a nonsignificant difference when data from the Greek rural sample were left out of the analysis ($t = .91$, $df = 308$, $p > .05$, $\eta^2 = .11$). The nonsignificant differences for the Greek total sample changed to significant differences in favor of the truncated Greek sample for grade 2 ($t = 2.18$, $df = 505$, $p < .05$, $\eta^2 = .20$) and grade 3 ($t = 2.72$, $df = 630$, $p < .05$, $\eta^2 = .24$). Rosenthal effect sizes confirmed that the differences between the full and truncated Greek sample and the American sample were negligible for all four grade levels ($\leq .30$). Because the effect sizes were small, the significant differences for all cross-cultural comparisons were likely attributable to the large sample sizes.

Split-halves reliability coefficients for the Tonal subtest in the Greek sample, although lower, are similar to those reported by Gordon (1979). The greatest difference was observed between the Greek ($r = .77$) and American ($r = .88$) kindergarten samples. Notably, reliability coefficients for the Rhythm subtest and composite test in the Greek sample were substantially lower than those Gordon reported for each grade level. Moreover, disparities among reliability coefficients for the Tonal, Rhythm, and composite means were far greater for the Greek sample than for the original norming sample. For example, with the Greek sample, the split-halves reliability coefficients for the composite test ranged from $r = .73$ to $r = .86$ for kindergarten through grade 3 compared to $r = .90$ to $r = .92$ for the U.S. sample (Table 1). Test–retest reliability coefficients were uniformly much lower for the Greek sample ($r = -.11$ to .64) than for the U.S. sample ($r = .60$ to .76). Gordon did not report the number of participants who took the PMMA twice, so perhaps his entire sample did so. In that case, the much smaller numbers of Greek participants involved in the retest portion of the study could explain some of the disparity in retest reliability between samples but probably not the entire disparity. Moreover, the test–retest results probably cannot be explained by developmental fluctuations in music aptitude in young children as Gordon (1979) claims, because grade 1

participants' scores were the most stable, whereas grade 3 scores were the least stable. Similarly, the relative lack of reliability in the Greek sample cannot be attributed to a lack of variance in the scores relative to the American sample (Table 1). Because the retest samples in the present study were relatively small, additional research is warranted.

The intercorrelations between the Tonal, Rhythm, and composite measures for the Greek sample were higher than those for the U.S. sample (Gordon, 1979) for each grade level, except for a slightly higher U.S. correlation coefficient for the Tonal and Rhythm subtests for the kindergarten samples and identical coefficients for the Rhythm subtest and composite test for grade 3. These higher correlations for the Greek sample may have resulted from the larger standard deviations in most subtests in the current study, likely attributable in part to the more heterogeneous Greek sample.

Mean item difficulty levels on the Tonal subtest for the Greek sample were slightly lower than those reported by Gordon (1979; kindergarten, .57 vs. .64; grade 1, .69 vs. .75; grade 2, .78 vs. .81; grade 3, .82 vs. .86, respectively). However, correlations for the difficulty levels between the Greek and U.S. samples were statistically significant and positive, ranging from *rho* = .75 to *rho* = .87 for grades 1 and 2, respectively (Table 4), indicating a moderately high degree of consistency in the difficulty level of the Tonal subtest items across the two samples. Item discrimination values were also similar on the Tonal subtest between the two samples, except for a lower value for grade 3 of the Greek sample compared to the U.S. sample (kindergarten, .36 vs. .37; grade 1, .33 vs. .38; grade 2, .30 vs. .38; grade 3, .24 vs. .36, respectively). All correlations for tonal item discrimination were also positive, and half were statistically significant (Table 4), but the coefficients were lower than those for tonal item difficulty because of the relatively restricted range of the item discrimination values. The results for both the Greek and U.S. samples suggest that as a whole, the Tonal subtest items have a modest level of discriminatory power. That is, the items do not tend to have great power to differentiate overall high versus low achievers on the tonal subtest.

Difficulty levels were nearly identical on the Rhythm subtest for the Greek and U.S. samples (kindergarten, .59 vs. .58; grade 1, .64 vs. .66; grade 2, .73 vs. .71; grade 3, .78 vs. .75, respectively). This similarity is reflected in the positive and statistically significant correlations for the Rhythm subtest between the two samples, ranging from moderate for grade 2 (*rho* = .39) to high for kindergarten (*rho* = .81) (Table 4). Greek discrimination values were somewhat lower than the U.S. sample for the Rhythm subtest (kindergarten, .25 vs. .28; grade 1, .25 vs. .31; grade 2, .28 vs. .33; grade 3, .23 vs. .34, respectively). As was the case for the Tonal subtest, correlations for rhythm item discrimination were lower than for rhythm item difficulty, with a range of *rho* = .38 to *rho* = .64. All correlations for item difficulty and item discrimination were positive between the Greek and U.S. samples, and most were statistically significant (Table 4).

Discussion

This study demonstrated the viability of a version of the PMMA adapted for use in Greece but with certain reservations. In most important respects, the Greek version of the test functioned as intended in that the overall results paralleled the American

Table 4. Spearman Correlations (*rho*) Between the Greek Standardization and Gordon (1979) Samples for Item Difficulty and Discrimination Values (*r*)

	Tonal		Rhythm	
Grade level	Difficulty	Discrimination	Difficulty	Discrimination
K	.81***	.28	.81***	.59***
1	.75***	.23	.53***	.64***
2	.87***	.36*	.39*	.42**
3	.78***	.53***	.49***	.38*

*$p < .05$. **$p < .01$. ***$p < .001$.

norms. Specifically, means by grade level were comparable to the American norms, higher group means were associated with higher grade levels, group means tended to be higher for the Tonal than for the Rhythm subtest, the internal consistency of the measure was adequate, and the observed trend for difficulty and discrimination values by grade level was as expected. Additionally, descriptive data for a sample of Greek children ages 5 and 6 years reported by Pollatou et al. (2005) are very much in line with the Greek norms determined for that age group in the current study.

Although the overall pattern of results was promising, a number of issues should be considered by test users and future researchers. The validity of the PMMA for the Greek population remains largely an open question, as the current methodology afforded a very limited assessment of validity. Although teachers' ratings of achievement correlated significantly with PMMA Tonal scores, they did not correlate with Rhythm scores. Furthermore, teachers' ratings of music aptitude were not correlated with PMMA subtests or the composite for any grade level. These results might be attributed to the noted low reliability of the Rhythm subtest in particular, possible lack of reliability in the teachers' ratings, or variation among teachers in their definitions and observations of children's music aptitude and achievement. Future research on the validity of the Greek version of the PMMA should include additional criterion measures of musical and nonmusical behaviors.

The differences noted between the Greek results and the original American norms can be viewed in light of general cultural differences, cultural differences in elementary music curricula, and the fact that the American norms were published in 1979. It is indeed open to question whether the original U.S. norms would be replicated in the United States 30 years later. As for the current Greek results, composite means (based on the total Greek sample) were significantly different from the American norms for grade 1, which suggests that test results in Greece should be interpreted according to the new Greek norms, not the original published norms. The somewhat larger standard deviations for the Greek sample should be considered also. Both the differences in means and especially the larger standard deviations can be explained by the fact that the Greek sample was obtained from a large cross-section of schools, classrooms, and school settings and thus was probably more heterogeneous than the original norming

sample, which consisted of a single suburban school district in the northeastern United States. Standard deviations for the Tonal and Rhythm subtests reported by Pollatou et al. (2005) for their Greek sample also exceeded the American norms. Statistical tests comparing the U.S. and Greek samples without the rural data revealed that the Greek urban and suburban composite means were significantly higher than the respective U.S. means for grades 2 and 3 but were not significantly different for kindergarten or grade 1.

Split-halves reliability coefficients for the Tonal subtest were similar for the Greek and American samples, except for notably lower coefficients for the kindergarten portion of the Greek sample. These coefficients, however, were considerably lower at all grade levels in the Greek sample for the Rhythm subtest, which had a similar effect on the composite coefficients. For reasons of reliability, the Tonal subtest of the Greek version of the PMMA developed for this study should be used with caution at the kindergarten level. The same can be said for the Rhythm subtest at the kindergarten and grade 1 levels.

Taken together, the internal and retest reliability and concurrent validity results for the Greek version of the PMMA are mixed. The coefficients for internal consistency for the Tonal and composite measures were within an acceptable range and were in line with those reported in previous research (e.g., Gordon, 1979; Gouzouasis, 1993; Holahan & Thomson, 1981). However, the internal consistency for the Rhythm subtest ranged from unacceptable to marginally acceptable and corroborated the results of Holahan and Thomson (1981). Internal reliability results for the Rhythm subtest should be viewed with caution, especially for kindergarten and grade 1 students. Similarly, as noted earlier, concurrent validity results based on teachers' ratings of students' music aptitude and achievement were poor for the Rhythm subtest at all grade levels.

It appears that Greek music educators and researchers can place more confidence in results from the Tonal than from the Rhythm subtest. However, the retest reliability results suggest that for all four grade levels, including older (i.e., grade 3) students, developmental music aptitude as measured by the Greek version of the PMMA was not stable across even a short period. The results also confirm that psychometric criteria for measurement instruments should not be assumed to be inherent qualities but, rather, should be continuously reassessed within every test sample.

Beyond the principal objective of cross-cultural issues examined here, the current findings suggest that future research should examine the stability of the PMMA, especially the Rhythm subtest. Although numerous studies, including those cited earlier, have reported results for internal consistency of the PMMA, relatively few have reported data concerning the stability of scores. Although some fluctuation in children's developmental music aptitude is to be expected (e.g., Gordon, 1979), the poor stability of scores in the current sample remains a major issue. As has been pointed out by Amos and Humes (1998), the lack of reliability concerns in widely used standardized measures of auditory perception (or any test instrument) may undermine the integrity of the interpretation of test scores and associated research conclusions and recommendations.

Attaining high retest reliability in tests of auditory perception with this age group does appear to be possible. For example, in a study of the reliability of the Slosson Auditory Perception Skills Screener, Erford and Luce (2005) found retest reliability of

r = .82 for a sample of 58 children ages 6 through 9. Notably, this 75-item measure was administered individually (in accordance with the test manual). This suggests that music education researchers should consider reexamining factors such as the administration condition, test length, and item characteristics that might influence the stability of the PMMA. Individual administration was used by Amos and Humes (1998) in their study of auditory discrimination in first and third graders. Interestingly, they found that 40-item auditory subtests (equal in length to the PMMA subtests) yielded poor retest reliability coefficients (r =.24 to r =.33), whereas 100-item and 180-item auditory subtests yielded retest reliability coefficients of .70 to .78, respectively.

Because the PMMA is a frequently used test, it is important that researchers continue to examine its reliability and validity over time and with different populations. The fact that three decades have passed since the first publication of the test points to the need for further investigation of its technical characteristics. Furthermore, the PMMA was intended to be used in the United States, and it was normed on a subsample of that population. Unfortunately, too often the test continues to be used uncritically with populations very different from the one it was intended to serve. Given the results of the current study, additional research on the reliability and validity of the PMMA, both within and between cultures, seems warranted.

Appendix

PMMA Tonal and Rhythm Percentile Norms for the Greek Sample

Tonal test					Rhythm test				
Raw score	Grade K	Grade 1	Grade 2	Grade 3	Raw score	Grade K	Grade 1	Grade 2	Grade 3
40				99	40				
39			99	98	39				
38			98	95	38				
37		99	96	88	37				99
36		98	92	79	36			99	96
35		96	85	68	35		99	96	89
34	99	93	73	57	34		98	92	77
33	98	83	61	44	33		96	84	66
32	97	74	50	35	32	99	93	75	56
31	96	68	39	26	31	97	90	61	45
30	94	63	33	21	30	93	86	55	33
29	87	55	26	16	29	89	79	44	28
28	85	53	21	13	28	86	70	36	22
27	82	46	17	12	27	83	62	29	18
26	75	40	14	10	26	78	52	23	14
25	70	36	12	9	25	70	44	19	11
24	58	32	10	7	24	63	39	15	9

Appendix (continued)

Tonal test					Rhythm test				
Raw score	Grade K	Grade 1	Grade 2	Grade 3	Raw score	Grade K	Grade 1	Grade 2	Grade 3
23	53	28	9	6	23	50	31	11	6
22	48	22	7	5	22	42	24	8	5
21	44	19	6	5	21	37	15	7	4
20	35	13	5	4	20	28	11	5	3
19	25	6	4	3	19	16	5	4	2
18	19	4	3	2	18	11	4	3	2
17	13	1	2	2	17	6	3	2	1
16	10		2	1	16	5	3	2	
15	8		1		15	3	2	1	
14	5				14	2	1		
13	3				13	1			
12	2				12				
11	1				11				

PMMA = Primary Measures of Music Audiation (Gordon, 1979).

Declaration of Conflicting Interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

Pythagoras II: Funding of research groups in the University of Macedonia, Priority Action 2.2.3e, Action 2.2.3, Measure 2.2, Operational Programme for Education and Initial Vocational Training, co-financed by the European Union (3rd Community Support Framework, 75% by the European Social Fund, 25% national Greek resources).

Note

1. In the absence of raw data from the original norming study, the authors computed a *t* test for each grade by hand after computing the respective variance coefficients for the U.S. samples (Gordon, 1979) via the following formula: $SD^2(N) = \Sigma x^2$. The Bonferroni adjustment for multiple tests was applied for each *t* test ($\infty/4$): $p < .05 = .0125$.

References

Amos, N. E., & Humes, L. E. (1998). SCAN test-retest reliability for first and third graders. *Journal of Speech, Language, and Hearing Research*, *41*, 834–845.

Bentley, A. (1966). *Measures of musical abilities.* London: Harrap Audio-Visual.

Chuang, W. C. J. (1997). *An investigation of the use of the Musical Aptitude Profile with Taiwanese students in grades four to twelve*. (Unpublished doctoral dissertation). Michigan State University, East Lansing.

Chung, H. (2002). *An investigation of the nature and characteristics of music aptitude of Korean-American students living in Pennsylvania*. (Unpublished doctoral dissertation). Temple University, Philadelphia.

Drake, R. (1933). Four new tests of musical talent. *Journal of Applied Psychology*, *17*, 136–147.

Drake, R. (1954). *Drake Musical Aptitude Tests.* Chicago: Science Research Associates.

Erford, B. T., & Luce, C. L. (2005). Reliability and validity of the Slosson Auditory Perception Skills Screener. *Perceptual and Motor Skills*, *101*, 891–897.

Feay-Shaw, S. (2000). Multicultural perspectives on research in music education. *Bulletin of the Council for Research in Music Education*, *145*, 15–26.

Gaston, E. T. (1957). *Tests of Musicality.* Lawrence, KS: Odell's Instrumental Service.

General Secretariat of National Statistical Service of Greece. (2008). De facto population: Area and population by range, density and elevation zones. Retrieved December 26, 2008, from http://www.statistics.gr

Gordon, E. E. (1965). *Musical Aptitude Profile.* Chicago: GIA.

Gordon, E. E. (1979). *Primary Measures of Music Audiation.* Chicago: GIA.

Gordon, E. E. (1982). *Intermediate Measures of Music Audiation.* Chicago: GIA.

Gordon, E. E. (1989). *Advanced Measures of Music Audiation.* Chicago: GIA.

Gordon, E. E. (1991). *The Advanced Measures of Music Audiation and the Instrument Timbre Test: Three research studies*. Chicago: GIA.

Gouzouasis, P. (1993). Music audiation: A comparison of the music abilities of kindergarten children of various ethnic backgrounds. *Quarterly Journal of Music Teaching and Learning*, *4*(2), 70–76.

Holahan, J. M., & Thomson, S. W. (1981). An investigation of the suitability of the Primary Measures of Music Audiation for use in England. *Psychology of Music*, *9*, 63–68.

Humphreys, J. T. (1993). Precursors of musical aptitude testing: From the Greeks through the work of Francis Galton. *Journal of Research in Music Education*, *41*, 315–327.

Humphreys, J. T. (1998). Musical aptitude testing: From James McKeen Cattell to Carl Emil Seashore. *Research Studies in Music Education*, *10*, 42–53.

Jung, K. H. (1992). *A two-year predictive validity study of the Musical Aptitude Profile for use in Korea*. Doctoral dissertation, Temple University, Philadelphia.

Kwalwasser, J. (1953). *Kwalwasser Music Talent Test.* New York: Mills Music.

Kwalwasser, J. (1955). *Exploring the musical mind.* New York: Coleman-Ross.

Kwalwasser, J., & Dykema, P. W. (1930). *Kwalwasser-Dykema Music Tests.* New York: Carl Fischer.

LeBlanc, A. (1980). Outline of a proposed model of sources of variation in musical taste. *Bulletin of the Council for Research in Music Education*, *61*, 29–34.

LeBlanc, A., Fung, C. V., Boal-Palheiros, G. M., Burt-Rider, A. J., Ogawa, Y., Oliviera, A. J., et al. (2002). Effect of strength of rhythmic beat on preferences of young music listeners in Brazil, Greece, Japan, Portugal, and the United States. *Bulletin of the Council for Research in Music Education*, *153*, 36–41.

LeBlanc, A., Jin, Y. C., Stamou, L., & McCrary, J. (1999). Effect of age, country, and gender on music listening preferences. *Bulletin of the Council for Research in Music Education*, *141*, 72–76.

Nunnally, J. C. (1970). *Introduction to psychological measurement*. Toronto, Ontario, Canada: McGraw-Hill.

Özeke, S., & Humphreys, J. T. (2000). Music teacher education in Turkey and in the USA: Musical aptitude and attitude toward teaching. In D. A. Tjeldvoll, T. Thune, & M. Sneve Olsen (Eds.), *Higher education, quality and evaluation in comparative and international perspectives* (Studies in Comparative and International Education Vol. 4, Report No. 7, pp. 67–80). Oslo, Norway: University of Oslo Institute for Educational Research.

Pollatou, E., Karidimou, K., & Gerodimos, V. (2005). Gender differences in musical aptitude, rhythmic ability, and motor performance in preschool children. *Early Childhood Development and Care*, *175*, 361–369.

Schoenoff, A. W. (1972). *An investigation of the comparability of American and German norms for the Musical Aptitude Profile.* (Unpublished doctoral dissertation). University of Iowa, Iowa City.

Seashore, C. E. (1919). *Seashore Measures of Musical Talents.* New York: Psychological Corporation.

Sell, V. H. (1976). *The musical aptitude of Finnish students: An investigative study in comparative music education*. (Unpublished doctoral dissertation). University of Wisconsin–Madison.

Stamou, L., Humphreys, J. T., & Schmidt, C. P. (2006). The effects of instruction on self-assessed research knowledge, ability, and interest among Greek music educators. *Music Education Research*, *8*, 175–189.

Tilson, L. (1941). A study of the prognostic value of the Tilson-Gretsch Test for Musical Aptitude. *Teachers College Journal*, *12*, 110–112.

Wang, J. C. (2007). *A comparative study of college students' musical aptitudes and musical preferences in the U.S. and Taiwan*. (Unpublished doctoral dissertation). Arizona State University, Tempe.

Whistler, H., & Thorpe, L. (1950). *Musical Aptitude Tests.* Los Angeles: California Test Bureau.

Wing, H. (1958). *Standardized Tests of Musical Intelligence.* Sheffield, UK: City of Sheffield Training College.

Yang, T. B. (2002). *The comparative effects of the traditional Taiwanese curriculum and a curriculum based on music learning theory on the developmental music aptitudes and singing performance of first grade students in Taiwan*. (Unpublished doctoral dissertation). Michigan State University, East Lansing.

Bios

Lelouda Stamou is an assistant professor of music education at the University of Macedonia in Thessaloniki, Greece. Her research interests include music aptitude development, study of musical behaviors, music and flow experience, and teacher training.

Charles P. Schmidt is a professor emeritus at Indiana University. His research interests include social psychology of music and instrumental music education.

Jere T. Humphreys is a professor of music education at Arizona State University. His research interests include historical and empirical aspects of scientific thinking in music education.

Submitted May 15, 2009; Acceptance October 22, 2009.