
Anthony J. Onwuegbuzie\textsuperscript{a} and Cindy L. Benge\textsuperscript{b}

\textsuperscript{a}Department of Educational Leadership, Sam Houston State University, TX, USA and Department of Educational Leadership and Management/Department of Educational Psychology, University of Johannesburg, Johannesburg, South Africa; \textsuperscript{b}Sam Houston State University, TX, USA

\textbf{ABSTRACT}
This Mixed Methods Theory-Based Impact Evaluation study focused on evaluating teacher implementation of the ARK (Adults Relating to Kids; Wilkerson, 1995/2013) program in secondary schools in an Independent School District (ISD). The researchers used the Classroom Assessment Scoring System—Adults Relating to Kids for Teachers and Staff Secondary Level (Onwuegbuzie & Benge, 2014) to conduct pre-ARK and post-ARK intervention observations of 10th-grade teachers at two experimental schools ($n = 44$) and one control school ($n = 18$). Quantitative and qualitative findings provided compelling evidence that the ARK intervention was mostly successful in increasing the effectiveness of the experimental school teachers.

\textbf{KEYWORDS}
Adults Relating to Kids Program; impact evaluation; mixed analysis; mixed methods; mixed methods analysis; Mixed Methods Theory-Based Impact Evaluation; mixed research; program evaluation

\textbf{Context and Purpose}

The present evaluation study focused on evaluating teacher implementation of the ARK (Adults Relating to Kids; Wilkerson, 1995/2013, 2007) program in secondary schools in an Independent School District (ISD) in a large state. This program evaluation, commissioned by the Rockwell Fund, Inc., was conducted in Year 4 of the projected 7-year project, which follows the cohort of students who entered Grade 7 in 2011 and who will graduate from high school in 2017. Although all teachers who teach students in the targeted ISD schools participate in ARK each year, this study focuses on the teachers of 10th-grade students during the 2014-2015 school year, students who would have been entering seventh grade in 2011, and, therefore, represent the original cohort of students targeted for ARK.

The original participating schools comprised all 10 middle school campuses and one alternative campus within the ISD. Although the ARK program is funded through various non-profit organizations, all campuses contribute monetarily to the project, forming a partnership with ARK and their funding partners. Five ninth-grade campuses and two high school campuses were added to the program during the 2013-2014 school year when the original cohort of students entered ninth grade. During the 2014-2015 school year, the original cohort of students entered 10th grade; therefore, all five senior high school campuses were added. Currently, all secondary schools in the ISD (i.e., 10 middle schools, five 9th-grade campuses, one alternative campus, and seven senior high schools) participate in ARK.

Longitudinal quantitative data collected throughout the first 4 years of the program include pre- and post-ARK surveys conducted in the fall and spring of each year, created to determine differences in students’ beliefs about teachers and about their own self-concepts before and after ARK teacher training each year. Additional quantitative data collected at the beginning and end of the year include student demographic data, discipline data, achievement data, and attendance data. Descriptive statistics are compiled and inferential statistical procedures are conducted to determine differences between beginning-of-year data and end-of-year data. An end-of-year teacher survey also is collected to gauge teacher attitudes and beliefs about ARK and their experience implementing ARK.
ARK

The ARKGroup, formerly The Children’s Center for Self-Esteem, is led by Dr. Glenn Wilkerson. The mission of the ARK organization is “to teach adults to give the unconditional love that builds high self-esteem and a positive self-concept in children” (Wilkerson, 2007, para. 1). ARK proposes to support a positive school climate that promotes (a) child and youth education, (b) high academic achievement, (c) dropout prevention, (d) violence prevention, (e) drug prevention education, (f) parent and community involvement, (g) teacher training and morale, (h) multicultural education, (i) health and wellness education, and (j) school innovation. To achieve these goals, the project director of ARK provides training to teacher representatives from each school throughout the school year to facilitate lessons targeting specific ways that teachers can reflect on and change behaviors that might negatively impact students' self-esteem. After the training, representatives then meet with a small group of teachers from their campuses two times per month for 1 hour. The meeting follows a three-item agenda: (a) review of the group covenants, (b) presentation of the lesson that includes a 5- to 10-minute video, and (c) a group-sharing period that includes situation circles (i.e., a group-sharing opportunity in which teachers share a situation and the group members listen and provide suggestions and feedback). After the meeting, teachers are expected to reflect on and to apply the lessons to their interactions with students. Teachers receive Continuing Professional Education (CPE) hours for attending each session.

Method

Instrument

For the evaluation of the ARK program, we developed the ARK Observation Protocol, which we named the Classroom Assessment Scoring System—Adults Relating to Kids for Teachers and Staff Secondary Level (CLASS-ARK-S; Onwuegbuzie & Benge, 2014) that incorporated all tenets of the ARK program. Utilizing the Secondary Classroom Assessment Scoring System (CLASS-S; Pianta, Hamre, & Mintz, 2012) as a model, the CLASS-ARK-S instrument involved the organization of the classroom expectations of ARK under the following three constructs utilized for the CLASS-S: (a) Emotional support, (b) Classroom organization, and (c) Instructional support (Pianta et al., 2012). For the CLASS-ARK-S, within each of these three constructs, three to six sub-constructs are included—yielding 13 subscales (i.e., sub-constructs) that comprise the three scales (i.e., constructs), as follows: (a) Emotional support, which comprises Positive Climate, Teacher Sensitivity, Regard for Adolescent Perspectives, and Nonverbal Interaction (i.e., 4 sub-constructs); (b) Classroom organization, which comprises Behavior Management, Productivity, and Negative Climate (i.e., 3 sub-constructs); and (c) Instructional support, which comprises Instructional Learning Formats, Content Understanding, Analysis and Inquiry, Quality of Feedback, Instructional Dialogue, and Student Engagement (i.e., 6 sub-constructs).

Further, bulleted items are provided within each sub-construct to specify key classroom concepts. Classroom observers then use the bulleted items of the CLASS-ARK-S to determine the level of implementation of the sub-constructs via a 4-point Likert-type scale (i.e., strongly agree, agree, disagree, strongly disagree). A 4-point scale was chosen in order to eliminate a midpoint (i.e., neutral) response option because inclusion of this response option has been found to reduce score reliability by more than one third (Weems & Onwuegbuzie, 2001), with nearly two thirds of respondents selecting the middle category more often than chance would predict—representing an over-reliance of this response option (Weems & Onwuegbuzie, 2001), and to place certain subgroups as being significantly more pre-disposed to select the mid-point response option (e.g., males; younger persons, persons with lower levels of academic-related self-perceptions, perfectionism, and/or hope; Onwuegbuzie & Weems, 2004).

The CLASS-ARK-S differed from Pianta et al.’s (2012) CLASS-S in three important ways. First, the CLASS-ARK-S includes an additional (i.e., 13th) dimension of Nonverbal Interaction because the ARK intervention emphasizes the importance of teachers using nonverbal communication cues such as appropriate physical touch, eye contact, focused attention, and active listening. Second, as noted previously, the CLASS-ARK-S comprises all the major tenets of the ARK intervention. Third, whereas the CLASS-S involved the use of a 7-point rating scale (i.e., low = 1-2, moderate = 3-5, and high = 6-7), observations pertaining to the CLASS-ARK-S were rated on a 4-point Likert-format scale. Appendix A presents the CLASS-ARK-S.

The CLASS-ARK-S was content-validated by the developer of the ARK intervention, namely Dr. B. Glenn Wilkerson, as well as by four teachers from the ISD who had been trained in and were extensively experienced in using the ARK intervention. Further, using scores obtained at the pre-intervention phase, the classical theory alpha reliability (i.e., α) for the total CLASS-ARK-S scale scores was .95 (95% confidence interval [CI] = .92, .96).
For the three sets of subscale scores, the classical theory alpha reliability was as follows: Emotional support (α = .89; 95% CI = .84, .93), Classroom organization (α = .78; 95% CI = .66, .86), and Instructional support (α = .91; 95% CI = .87, .94). These alpha values indicate that the total scale and three subscales yielded scores with high reliability.

Procedures

During the Fall 2014 semester, beginning on October 6, 2014, we used the CLASS-ARK-S to conduct a series of pre-ARK intervention observations of the ISD 10th-grade teachers at two experimental schools (i.e., Experimental High School 1 and Experimental High School 2). All the observations at these two schools took place in October, except for one interview, which took place on November 3, 2014 due to scheduling conflicts. These experimental school teachers were the current teachers of members of the original cohort of students (i.e., members of the graduating class of 2017). Further, between December 9, 2014 and December 16, 2014, one of the evaluators conducted a series of pre-ARK observations of 10th-grade teachers at the control school.

To assess the inter-rater reliability between the two observers (i.e., the two authors of this evaluation report), the initial video recordings of two classes were used. Specifically, we independently watched the video tape of the pre-Ark intervention first lesson and rated this first lesson using the CLASS-ARK-S. These two sets of independent ratings of the pre-Ark intervention first lesson then were compared and any discrepancies were discussed. These discrepancies led to the development of scoring criteria. These criteria include scoring rules for when teachers (a) spent an inordinate time lecturing, (b) ended the class with a worksheet (i.e., homework), and (c) did not give students an opportunity to work together in groups to work through a problem. Next, we independently watched the video tape of the pre-Ark intervention second lesson and rated this second lesson using the CLASS-ARK-S. These two sets of independent ratings of the pre-Ark intervention second lesson yielded an inter-rater reliability of 100%. As a result, the remaining 60 observations were divided between the two raters, and were coded individually.

During the Spring 2015 semester, we conducted a series of post-ARK intervention observations of the same set of pre-ARK experimental school teachers over a four-week period, beginning on April 21, 2015 and ending on May 18, 2015. During this same period, one of the evaluators conducted a series of post-ARK observations of the same set of pre-ARK control school teachers, beginning on May 13, 2015 and ending on May 15, 2015.

Campus administrators were notified in advance via email that we would be conducting observations and were given the opportunity to reschedule observations if necessary. Teachers were not notified in advance in order to minimize any threat to internal validity stemming from reactive arrangements (i.e., Hawthorne effect, John Henry effect, resentful demoralization, the novelty effect, and placebo effect; Campbell, 1957; Campbell & Stanley, 1963; Cook & Campbell, 1979; Onwuegbuzie, 2003; Shadish, Cook, & Campbell, 2001). However, in the vast majority of cases, we conducted our post-ARK observations on the same day of the week and at the same time period as occurred during the pre-ARK observations. In doing so, our goal was to minimize any threats to the internal validity of the findings relating to a time x treatment interaction (Campbell, 1957; Campbell & Stanley, 1963; Cook & Campbell, 1979; Onwuegbuzie, 2003; Shadish et al., 2001).

Participants

Originally, before our evaluation study begun, we had planned to observe between 25 and 30 teachers in both of the experimental schools and the control schools for a total of between 75 and 90 observations. However, we made this projection before we knew how many teachers were available for observation at each of the three schools. We ended up with a total sample size of 62, comprising 44 teachers from the experimental schools (i.e., 18 teachers from Experimental High School 2 and 26 teachers from Experimental High School 1) and 18 teachers from the control school. These numbers represent all the teachers assigned to us by the campus administrators who were available during our periods of observation. Notwithstanding, the total sample size of 62 provided adequate statistical power (i.e., .99) to detect a statistically significant difference among the three schools across the three CLASS-ARK-S scale scores via a multiple analysis of variance (MANOVA) at the 5% level of statistical significance and a moderate effect size (i.e., $f = .25$). Further, this total sample size of 62 provided adequate statistical power (i.e., .86) to detect a statistical significant difference in post-ARK minus Pre-ARK differences among the three schools across the 13 CLASS-ARK-S subscale scores via a multiple analysis of variance (MANOVA) at the 5% level of statistical significance and a moderate effect size (i.e., $f = .25$). Thus, the sample size of 62 was more than adequate. Indeed, this sample size yielded a total of 4,860 minutes of observation. Moreover, the 62 classrooms that we observed far exceeded the 49 kindergarten classroom teachers from 49
kindergarten classrooms examined by Pakarinen et al. (2010) using the CLASS Pre-K (Pianta, La Paro, & Hamre, 2008), as well as the 49 kindergarten classroom teachers from 47 preschool classrooms from 23 schools examined by Downer, Booren, Lima, Luckner, and Pianta (2010) using the Individualized Classroom Assessment Scoring System (inCLASS). Of the sample of 62 teachers, the slight majority (53.2%) were women.

We observed a wide range of classes in the three schools wherein the following 26 subjects were taught: algebra, Advance Placement (AP) chemistry, art, band, biology, chemistry, criminal justice, dance, English language arts (ELA), ELA honors, engineering, English, English as a second language (ESL), geometry, AP government, history, integrated physics and chemistry, mathematics, pre-AP chemistry, professional communication, reading, Reserve Officers’ Training Corps (ROTC), science/physics, social studies, and world history. The size of the classes observed ranged from four to 30 ($M = 17.38, SD = 6.24; N = 869$), with the number of boys in these classes ranging from 0 to 24 ($M = 9.44, SD = 4.64; n = 472$) and number of girls in these classes also ranging from 0 to 24 ($M = 7.94, SD = 4.42; n = 397$). An independent samples t-test revealed that, overall, these classrooms contained statistically significantly more boys than girls ($t(867) = 4.85, p < .0001$), yielding a small-to-moderate Cohen’s (1988) d effect size of 0.33.

**Evaluation Design**

The present evaluation was part of a larger 7-year program evaluation of the ARK intervention. This 7-year evaluation of ARK has been undertaken using an eclectic approach, as recommended by Fitzpatrick, Sanders, and Worthen (2011). Specifically, this evaluation has involved combining the following two broad evaluation approaches: Decision-Oriented Evaluation and Program-Oriented Evaluation.

**Decision-Oriented Evaluation.** The decision-oriented evaluation approach has involved use of the Context, Input, Process and Product (CIPP) Evaluation Model (Stufflebeam, 1971, 2004, 2005, 2007). According to Stufflebeam (2005), CIPP evaluation is “the process of delineating, obtaining, reporting and applying descriptive and judgmental information about some object’s merit, worth, probity, and significance to guide decision making, support accountability, disseminate effective practices, and increase understanding of the involved phenomena” (p. 61). The CIPP Evaluation model involves the evaluation of the following four different types of decisions: (a) Context evaluation, which involves planning decisions by studying the context for a program by determining what needs are to be addressed by the program and what programs currently exist as a means of defining/redefining the program; (b) Input evaluation, which involves structuring decisions by selecting a specific strategy to implement and to resolve problems, and making decisions about how to implement it; (c) Process evaluation, which involves implementing decisions by deciding how to modify its implementation as needed; and (d) Product evaluation, which involves assessing decisions by assessing both intended and unintended outcomes. Table 1 shows the primary components of the CIPP, as proposed by Stufflebeam (2005) and Stufflebeam and Shinkfield (2007).

<table>
<thead>
<tr>
<th>Evaluation roles</th>
<th>Context</th>
<th>Input</th>
<th>Process</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Formative evaluation:</strong></td>
<td>Guidance for identifying need for ARK and for selecting and making goals/objectives based on assessing needs, problems, assets, opportunities, and guiding questions for ARK training and program experiences</td>
<td>Guidance for choosing ARK (based on assessing alternative strategies and resource allocation plans) and for examining work plan</td>
<td>Guidance for implementing the ARK plan (based on monitoring and assessing program activities)</td>
<td>Guidance for continuing, modifying, adopting, or terminating ARK (based on assessing the intended and unintended outcomes)</td>
</tr>
<tr>
<td><strong>Prospective application of CIPP information to facilitate decision making and quality control</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Summative evaluation:</strong></td>
<td>Comparison goals and objectives to assess needs, problems, assets, and opportunities</td>
<td>Comparison of ARK’s strategy, design, and budget to other programs and the identified needs of stakeholders</td>
<td>Full description of the ARK process (cf. Wilkerson, 1995/2013) and costs, as well as comparison of the designed and actual processes and costs</td>
<td>Comparison of intended and unintended outcomes to targeted needs and, where possible, to the results of other programs; interpretation of results against the ARK’s assessed context, inputs, and processes</td>
</tr>
<tr>
<td><strong>Retrospective use of CIPP information to describe and to assess the program’s merit, worth, probity, and significance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Program-Oriented Evaluation.** Alongside the decision-oriented CIPP Evaluation Model, a program-oriented evaluation model is being used. Specifically, a Theory-Driven Evaluation Model (Chen, 1990; Chen & Rossi, 1980, 1983) has been employed. Chen (2006) stated that “In theory-driven evaluations, the application of mixed methods is justified and applied under a conceptual framework called program theory” (p. 75). Donaldson (2007) defined program theory as “the process through which program components are presumed to affect outcomes and the conditions under which these processes are believed to operate” (p. 22). As recommended by Chen (1990), developing the program theory involves rigorously searching the scientific research literature to identify social science theories (e.g., theories related to culturally responsive leaders). Developing the program theory also involves input from stakeholders (Donaldson, 2007). As recommended by Donaldson (2007), the Theory-Driven Evaluation Model used to date has involved the following steps: (a) engage relevant stakeholders (e.g., teachers, principals); (b) develop an initial draft of the program theory; (c) present the draft to the stakeholders for feedback; (d) conduct a plausibility audit (i.e., assess the plausibility of the link between the program action and intended outcome); (e) communicate these findings to key stakeholders and revise the program theory as needed; (f) probe links for model specificity (i.e., examining the program theory at a deeper level); and (g) finalize program impact theory. Table 1 presents all instruments/metrics that have been/will be used.

The current evaluation was what could be called an impact evaluation. Onwuegbuzie and Hitchcock (2017) recently provided a new comprehensive definition of impact evaluation as follows:

Impact evaluation is the rigorous and systematic analysis conducted collaboratively by project staff, evaluators, and local stakeholders of the long-term changes—positive or negative and intended or unintended—in the lives of a person, group, or community that stem from an observable set of actions relative to a credible counterfactual that includes an analysis of the process underlying the intervention, treatment, program, or policy, with the ultimate goal of the local stakeholders using the information about the process and outcome extracted from this analysis to improve their situation. (p. 58)

Moreover, the present program evaluation involved the use of both qualitative and quantitative approaches, yielding what Onwuegbuzie and Hitchcock (2017) conceptualized as a *Mixed Methods Theory-Based Impact Evaluation*, which involved the incorporation into a mixed methods-based impact evaluation of some kind of theory, whether evaluation theory (i.e., representing guiding criteria that indicate what an appropriate evaluation is and how evaluation should be conducted), social science theory (i.e., representing a framework for understanding the nature and etiology of desired or undesired outcomes and for developing intervention strategies for influencing those outcomes), and/or program theory (i.e., focusing on the assumptions that underlie the specific interventions/treatment/programs and how they are expected to bring about change). Onwuegbuzie and Hitchcock (2017) defined a MMTBIE, as follows:

A Mixed Methods Theory-Based Impact Evaluation is a rigorous and systematic approach, framed by evaluation theory, social science theory, and/or program theory, involving the use of quantitative and qualitative evaluation techniques that is conducted collaboratively by project staff, evaluators, and local stakeholders to assess in an informative, complete, and balanced way the long-term changes—positive or negative and intended or unintended—in the lives of a person, group, or community that stem from an observable set of actions relative to a credible counterfactual. This form of impact analysis relies on qualitative and quantitative viewpoints, data collection, analysis, and inference techniques combined according to the logic of mixed methods evaluation. Such evaluation also monitors the process underlying the intervention, treatment, program, or policy, with the ultimate goal of the local stakeholders using the information about the process and outcome extracted from this analysis to improve their situation. (p. 58)

The form of MMTBIE that we used in the current investigation involved the use of social science theories such as Dweck’s (1986) theory of motivation and Goleman’s (1995, 2005) theory of emotional intelligence. According to Onwuegbuzie and Hitchcock (2017), the MMTBIE involves the following eight phases: (a) Phase 1: understand the local and broader context; (b) Phase 2: understand the construct(s) of interest; (c) Phase 3: map out the causal chain that explains how the intervention is expected to produce the intended outcomes; (d) Phase 4: collect quantitative and qualitative data to test the underlying assumptions of the causal links; (e) Phase 5: determine the type and level of generalizability and transferability; (f) Phase 6: conduct a rigorous evaluation of impact; (g) Phase 7: conduct a rigorous process analysis of links in the causal chain; and (h) Phase 8: conduct a meta-evaluation of the process and product of the MMTBIE. In this evaluation, we focus on Phases 4-8 because information pertaining to Phases 1-3 are explained elsewhere (e.g., Wilkerson, 1995/2013, 2007).

Impact evaluations can be treated as lying on a continuum, with efficacy studies and effectiveness studies lying on opposite ends. Broadly speaking, efficacy can be defined as the performance of a particular intervention, treatment, program, or policy under ideal and controlled conditions or circumstances (Singal, Higgins, & Waljee, 2014). In contrast, effectiveness refers to its performance under real-world conditions (Singal et al., 2014). The
present program evaluation represented an effectiveness study because it involved an evaluation of the extent to which teachers who underwent the ARK training incorporated its tenets into their day-to-day teachings.

With regard to Phase 4 of the MMTBIE, we used what Leech and Onwuegbuzie (2009) referred to as a fully mixed concurrent dominant status design, which involves conducting a study that mixes qualitative and quantitative research within one or more of, or across the stages of the research process. In the MMTBIE, the quantitative and qualitative phases occurred concurrently at one or more stages or across the stages, with the qualitative phase and quantitative phase being given approximately equal weight.

With respect to Phases 6 and 7, conducting a rigorous evaluation of impact and conducting a rigorous process analysis of links in the causal chain, respectively, we used a credible counterfactual, by measuring what would have happened to beneficiaries of the intervention (i.e., teachers) in its absence, with the impact being estimated by comparing counterfactual outcomes to those observed under the intervention. Specifically, we defined the counterfactual in terms of a control group. Importantly, our use of a quasi-experimental approach was combined with the use of qualitative approaches to help rule out rival hypotheses and to determine more reliably the causal chain (Onwuegbuzie & Hitchcock, 2017)—yielding an integrated approach to impact evaluation (Onwuegbuzie, 2017). In so doing, we operated under the assumption that the effect—that is, teachers being able to relate to their students effectively—can be attributed directly to the ARK intervention.

Results

Univariate Analysis of Overall CLASS-ARK-S Scale Scores

The total CLASS-ARK-S Scale scores revealed that whereas the overall teaching effectiveness for the control school reduced from the pre-ARK intervention period to the post-ARK intervention period (i.e., \( M = -1.59, SD = 4.85 \)), both experimental schools—namely Experimental High School 1 (i.e., \( M = 2.43, SD = 4.83 \)) and Experimental High School 2 (i.e., \( M = 2.00, SD = 5.83 \))—experienced gains from the pre-ARK intervention period to the post-ARK intervention period. Using Onwuegbuzie and Daniel’s (2002) criteria for a standardized skewness coefficient (i.e., skewness coefficient divided by its standard error) and a standardized kurtosis coefficient (i.e., kurtosis coefficient divided by its standard error), wherein standardized skewness coefficients and standardized kurtosis coefficients that lie outside the ±3 range indicate serious departures from normality, the difference scores (i.e., post-ARK intervention – pre-ARK intervention) pertaining to the total CLASS-ARK-S Scale (i.e., standardized skewness coefficient = 3.12, standardized kurtosis coefficient = 6.09) suggested a serious departure from normality. Thus, a nonparametric analysis of variance (ANOVA) was used to compare the difference in scores among these three schools. Specifically, the Kruskal-Wallis test was employed.

The nonparametric ANOVA revealed a statistically significant difference among the three schools, \( \chi^2(2) = 7.59, p = .02 \). The effect size associated with this difference, as measured by Cramer’s \( V \), was 0.25. Using Cohen’s (1988) criteria, this coefficient indicated a moderate effect. A series of nonparametric pairwise follow-up tests (i.e., Mann-Whitney’s \( U \)) was conducted to examine further the nature of the differences among the three schools. The Bonferroni adjustment was applied to take into account the fact that three pairwise follow-up tests were undertaken, such that the total experimentwise error rate did not exceed 5% (Chandler, 1995; Ho, 2006; Manly, 2004; Vogt, 2005). This correction was undertaken by dividing the nominal alpha value by three (i.e., \( .05/3 = .0167 \)). Therefore, the adjusted level of statistical significance was .0167. After applying the Bonferroni adjustment, the Mann-Whitney’s \( U \) tests indicated that Experimental High School 1 had statistically significantly higher gains than did the control school. In contrast, although Experimental High School 2 had higher gains than did the control school, this difference was not statistically significant (\( U = 75.50, p = .05 \)). Similarly, although Experimental High School 1 had higher gains than did Experimental High School 2, this difference was not statistically significant (\( U = 122.00, p = .25 \)).

CLASS-ARK-S Scale Scores and Subscale Scores

Table 2 presents descriptive statistics pertaining to the 17 sets of Post-ARK-Pre-ARK differences as a function of school: the CLASS-ARK-S overall score, three scale scores (i.e., Emotional support, Classroom organization, and Instructional support), and 13 subscale scores. From this table, it can be seen that, of these 17 indices, 15 of them involved greater gains for both the experimental schools than for the control school. Indeed, only the subscales Analysis and Inquiry and Instructional Dialogue yielded differences in favor of the control school.
Onwuegbuzie and Levin (2005) conceptualized three methods for testing the trend of multiple group differences, with the Binomial Test of Trend being perhaps the most notable, in which the binomial distribution is used to determine whether the number of group differences in the same direction (positive/negative) should be regarded as either statistically real or representing a statistically chance finding. Thus, using this procedure, the probability that all 15 or more differences out of 17 are in favor of the experimental schools is .0012, which indicates that the trend was statistically significant at the .05 level, with a large effect size of 0.88 (i.e., 15/17). Using Onwuegbuzie, Levin, and Ferron’s (2011, p. 130) table that specifies the number of directional group differences required for statistical significance (α = .05, one-tailed) as a function of the number of measures examined (i.e., 17) and the correlation between measures (not presented), this conclusion is valid even if the intermeasure correlation had been as high as .20, which was not the case. Thus, the Binomial Test of Trend provides even more compelling evidence of the effectiveness of the ARK intervention.

### Multivariate Analyses of CLASS-ARK-S Scale Scores

In order to determine further the nature of the differences in CLASS-ARK-S scale scores between the two experimental schools and the control school, a one-way MANOVA was conducted to examine simultaneously differences among the three schools with respect to the three scale scores (i.e., Emotional support, Classroom organization, and Instructional support). Because all three sets of scale scores departed from normality—each indicating positive skew with a leptokurtic distribution that was characterized by a shape that was more peaked—caution should be exercised in interpreting results stemming from the multivariate analysis of variance (MANOVA). Also, it should be noted that MANOVA requires the assumption of multivariate normality, which is more restrictive than the univariate normality assumed by its univariate counterpart, namely, analysis of variance. The assumption of multivariate normality is more stringent because it implies not only that each individual variable has a normal distribution but also that the joint distribution of the variables is multivariate normal (Bray & Maxwell, 1985). Thus, although univariate normality is a necessary condition for multivariate normality, it is not a sufficient one. That is, univariate normality “does not guarantee multivariate normality” (Field, 2009, p. 604, italics in original).

Prior to conducting the MANOVA, Box’s M test was conducted to assess the homogeneity of the variance-covariance matrix involving the four variables of interest (Tabachnick & Fidell, 2007). Box’s M statistic was statistically significant at 36.87, which suggested heterogeneity of the variance-covariance matrix \( F[12, 10200.96] = 2.80, p < .001 \). Because of the presence of heterogeneity, caution should be exercised in interpreting the results of the MANOVA.

The MANOVA revealed a statistically significant main effect for school \( F[3, 49] = 4.21, p = .01 \). Further, the effect size, as measured by partial \( \eta^2 \), associated with this school difference was 0.21. Using Cohen’s (1988)
criteria, this coefficient indicates a large effect. Thus, there were statistically significant and practically significant differences among the schools with respect to the three scale scores simultaneously. Because the MANOVA yielded a statistically significant main effect, follow-up discriminant analyses were conducted, as recommended by many statisticians (e.g., Field, 2009; Huberty & Morris, 1989; Keselman et al., 1998; Onwuegbuzie & Daniel, 2003). Specifically, two discriminant analyses were conducted—the first discriminant analysis involving the comparison of Experimental High School 1 and the control school and the second discriminant analysis involving the comparison of Experimental High School 2 and the control school.

**Experimental High School 1 versus Control School.** The discriminant analysis revealed a statistically significant canonical function ($\eta_1^2 = 0.49$, $p < .024$; Wilks’s Lambda = 0.76). The corresponding canonical correlation was .49, which suggested a large effect size (Cohen, 1988). In addition, the group centroid (the average score on the discriminant function for both schools) for this function was .49 for Experimental High School 1 and -.61 for the control school. These statistics indicated that the discriminant function maximally separated these two schools.

An examination of the standardized canonical discriminant function coefficients (Table 3) revealed that, using a cutoff loading of 0.3 (Lambert & Durand, 1975; Tabachnick & Fidell, 2007), emotional support and classroom organization scores were practically significant. Further, the structure coefficients (i.e., structure matrix) between the independent variable set and the standardized canonical discriminant function (Table 3) indicated that, using a cutoff loading of 0.3 (Lambert & Durand, 1975; Tabachnick & Fidell, 2007), all three variables discriminated the two schools. Classroom organization score was the most significant predictor of school, followed by emotional support. Variables with a positive coefficient (i.e., classroom organization, emotional support) suggest that teachers high on these variables were more likely to be classified as belonging to Experimental High School 1. Conversely, a variable with a negative coefficient (i.e., instructional support) indicates that teachers high on this variable were more likely to be categorized as belonging to the control school. A comparison of the standardized and structure coefficients revealed no suppressor variables because all the variables with significant standardized coefficients (i.e., $>.30$) also had significant structure coefficients (Henson, 2002; Onwuegbuzie & Daniel, 2003). However, the fact that instructional support had a significant structure coefficient but a non-significant standardized coefficient suggests that this variable might have been collinear (Onwuegbuzie & Daniel, 2003).

**Experimental High School 2 versus the Control School.** The discriminant analysis did not reveal a statistically significant canonical function ($\eta_2^2 = 0.62$, $p = .10$; Wilks’s Lambda = 0.80). As such, no further analysis was undertaken.

**Multivariate Analyses of CLASS-ARK-S Subscale Scores**

In order to determine further the nature of the differences in CLASS-ARK-S subscale scores between the two experimental schools and the control school, a one-way MANOVA was conducted to examine simultaneously differences among the three schools with respect to the 13 subscale scores. Because all but one of these variables (i.e., behavior management) departed from normality—some indicating a positive skew and some indicating a negative skew, all with a leptokurtic distribution that was characterized by a shape that was more peaked—caution should be exercised in interpreting results stemming from the multivariate analysis of variance (MANOVA).

The Box’s $M$ statistic was statistically significant at 36.87, which suggested heterogeneity of the variance-covariance matrix ($F_{13,39} = 2.29, p = .02$). Because of the presence of heterogeneity, caution should be exercised in interpreting the results of the MANOVA. The MANOVA revealed a statistically significant main effect for school ($F_{13,39} = 2.29, p = .02$). Further, the effect size, as measured by partial $\eta^2$, associated with this school difference was 0.43. Using Cohen’s (1988) criteria, this coefficient indicates a very large effect. Thus, there were statistically significant and practically significant differences among the schools with respect to the 13 subscale scores simultaneously. Because the MANOVA yielded a statistically significant main effect, two follow-
up discriminant analyses were conducted—the first discriminant analysis involving the comparison of Experimental High School 1 and the control school and the second discriminant analysis involving the comparison of Experimental High School 2 and the control school.

**Experimental High School 1 versus the Control School.** The discriminant analysis revealed a statistically significant canonical function ($\eta^2[13] = 21.00, p < .05$; Wilk’s Lambda = 0.49). The corresponding canonical correlation was .71, which suggested an extremely large effect size (Cohen, 1988). In addition, the group centroid for this function was 0.89 for Experimental High School 1 and -1.10 for the control school. These statistics indicated that the discriminant function maximally separated these two schools.

An examination of the standardized canonical discriminant function coefficients (Table 4) revealed that, using a cutoff loading of 0.3 (Lambert & Durand, 1975; Tabachnick & Fidell, 2007), nonverbal interaction, negative climate, instructional dialogue, analysis and inquiry, and instructional learning formats were practically significant, with nonverbal interaction being the most significant predictor of school. Further, the structure coefficients (Table 4) indicated that nonverbal interaction, regard for adolescent perspectives, behavior management, negative climate, and student engagement discriminated the two schools, with nonverbal interaction having the largest effect size. Variables with a positive coefficient (i.e., nonverbal interaction, negative climate, instructional learning format, regard for adolescent perspectives, behavior management, and student engagement) suggest that teachers high on these variables were more likely to be classified as belonging to Experimental High School 1. Conversely, variables with a negative coefficient (i.e., analysis and inquiry and instructional dialogue) indicate that teachers high on this variable were more likely to be categorized as belonging to the control school. A comparison of the standardized and structure coefficients revealed that instructional learning formats, analysis and inquiry, and instructional dialogue served as suppressor variables because they were all variables with significant standardized coefficients (i.e., > .30) but with non-significant structure coefficients (Henson, 2002; Onwuegbuzie & Daniel, 2003). Further, the fact that regard for adolescent perspective behavior management and student engagement had significant structure coefficients but non-significant standardized coefficients suggests that these variables might have been collinear (Onwuegbuzie & Daniel, 2003). In summary, based on their significant standardized coefficients and structure coefficients, nonverbal interaction and negative climate played the largest role in discriminating the two schools.

**Table 4. Standardized and Structure Coefficients for the 13 Subscale Scores: Experimental High School 1 versus Control School**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Standardized Coefficient</th>
<th>Structure Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Climate</td>
<td>-.17</td>
<td>.27</td>
</tr>
<tr>
<td>Teacher Sensitivity</td>
<td>-.20</td>
<td>.20</td>
</tr>
<tr>
<td>Regard for Adolescent Perspectives</td>
<td>.08</td>
<td>.39*</td>
</tr>
<tr>
<td>Nonverbal Interaction</td>
<td>.63*</td>
<td>.52*</td>
</tr>
<tr>
<td>Behavior Management</td>
<td>.11</td>
<td>.39*</td>
</tr>
<tr>
<td>Productivity</td>
<td>.14</td>
<td>.23</td>
</tr>
<tr>
<td>Negative Climate</td>
<td>.51*</td>
<td>.53*</td>
</tr>
<tr>
<td>Instructional Learning Formats</td>
<td>.43*</td>
<td>.24</td>
</tr>
<tr>
<td>Content Understanding</td>
<td>-.14</td>
<td>.12</td>
</tr>
<tr>
<td>Analysis and Inquiry</td>
<td>-.48*</td>
<td>-.13</td>
</tr>
<tr>
<td>Quality of Feedback</td>
<td>-.02</td>
<td>.15</td>
</tr>
<tr>
<td>Instructional Dialogue</td>
<td>-.51*</td>
<td>-.20</td>
</tr>
<tr>
<td>Student Engagement</td>
<td>.29</td>
<td>.49*</td>
</tr>
</tbody>
</table>

Note: *Coefficients with effect sizes larger than .3 (Lambert & Durand, 1975).

**Experimental High School 2 versus the Control School.** The discriminant analysis did not reveal a statistically significant canonical function ($\eta^2[13] = 21.41, p = .07$; Wilk's Lambda = 0.40). As such, no further analysis was undertaken.

**Qualitative Findings**

A total of 34 of the 44 teachers in the experimental schools (77.2%) undertook a brief interview conducted by the observer after the class ended. Overall, the vast majority of these teachers ($n = 30$; 88.2%) were positive about the ARK program, representing a large effect size. A constant comparison analysis (Glaser, 1965) of the interview responses yielded seven major themes. Whereas three of these themes were negative in nature, the remaining four were positive.
The negative themes were labeled as No value added, Affirmations, and ARK Video. These themes most emerged from the more experienced teachers. Exemplar statements pertaining to the No value added theme made by the teachers included the following:

Affirmations were just commonsense.
It’s just normal teaching.
ARK did not teach me anything new.

In contrast, the Affirmations theme, the second negative theme, stemmed from criticism of a few teachers regarding the ARK teacher covenants—especially the self-affirmations, which, according to the ARK manual, are as follows:
1) I am a unique and lovable person.
2) I am a person of tremendous worth.
3) Although I make mistakes, I am worthy of forgiveness.
4) I deserve to be loved for who I am rather than for what I do.
5) I accept and love myself.

The third negative theme, ARK video, stemmed from criticisms that a few teachers made about the ARK video. One person declared the following: Person on the ARK video was not enthusiastic, which turned me off.

On the positive side, the most prevalent theme of all seven themes was labeled Good Idea, wherein most of the teachers acknowledged the potential of ARK. A second positive theme that emerged was labeled Building Relationships. As an example of this theme, an African American female reading teacher at one of the experimental schools declared the following:

I think that ARK is about building relationships ... not to get into power struggles ... to have a reminder they are kids ... admonishment with love ... the training just needs to be sooner during the year.

The third theme that emerged was labeled Refresher, wherein some of the teachers—usually the ones who had been in the teaching profession for a relatively long time—concluded that the ARK program provided them with an opportunity to review some of the best teaching practices, as exemplified by the following statement: “ARK was a good refresher.”

The fourth emergent theme was labeled De-escalation. This theme provided one of the most compelling qualitative examples of the effectiveness of the ARK program, wherein during the interview, an African American chemistry teacher at one of the experimental schools stated that “The ARK program taught me how to de-escalate.” Interestingly, just prior to this interview, the observer of this teacher actually documented actions that he made in class that were consistent with this statement. In particular, during his chemistry class, when the most disruptive student was disrespectful to the teacher, he de-escalated the situation by ignoring what the student stated, displaying even more respect for the student by positively affirming him, and re-directing the student back to the in-class assignment—all tenets of the ARK intervention.

On balance, the positive themes that emerged were significantly more prevalent than were the negative themes. However, the negative themes provide some direction for improvement of the ARK project, some of which we present in the next section. Interestingly, the teachers who had been in the teaching profession for a shorter period of time tended to be more positive about the ARK intervention than were their counterparts who had more teaching experience.

Discussion

The present evaluation is unique in at least four ways. First, it is the first evaluation in which mixed methods research techniques were used to evaluate the ARK intervention, namely, a MMTBIE. Second, with nearly 5,000 minutes of observation, this impact evaluation is more comprehensive than the vast majority of impact evaluations that involve classroom observations. Third, this evaluation also is comprehensive inasmuch as it involved the observation of the effectiveness of ARK project across 26 subject areas. Fourth, the present evaluation involved the video recording of every single lesson that was observed, which helped to maximize the score reliability and score validity of the observation ratings, as exemplified by the large score reliability coefficients for the CLASS-ARK-S total scale and three subscales. By designing a rigorous and comprehensive impact evaluation, we were able to examine thoroughly the effectiveness of the ARK intervention. More specifically, by using a counterfactual analysis, we have been able to rule out confounding variables before attributing the observed changes in indicators to the actual ARK intervention.
With these considerations in mind, the quantitative and qualitative findings provide compelling evidence that the ARK intervention thus far has been mostly successful in increasing the effectiveness of the select teachers at the two experimental schools. Specifically, at Experimental High School 2, the ARK program was found to be at least partially successful inasmuch as teachers from this school, in general, had positive gains in their teacher effectiveness, as opposed to teachers in the control school, who, experienced deficits—although the overall post-pre difference between these two groups was not statistically significant. Also, as a set, the three ARK constructs did not statistically significantly discriminate the two schools, nor did the 13 ARK subcontracts. However, it should be noted that 15 of the 17 differences—comprising all three ARK constructs and 12 of the 14 ARK subconstructs—that were measured between these two schools were in favor of Experimental High School 2, which was both statistically significant with a large effect size of 0.88. As such, the gains in teacher effectiveness of the select teachers at Experimental High School 2 compared to the control school are notable.

Even more notable was the finding pertaining to Experimental High School 1. Specifically, (a) the overall post-pre difference between Experimental High School 1 and the control school was both statistically significant and practically significant; (b) a total of 15 of the 17 differences—comprising all three ARK constructs and 12 of the 14 ARK subconstructs—that were measured between these two schools were in favor of Experimental High School 1, which was both statistically significant with a large effect size of 0.88; (c) as a set, the three ARK constructs statistically significantly discriminated the two schools in favor of Experimental High School 1; and (d) as a set, the 13 ARK subcontracts statistically significantly discriminated the two schools in favor of Experimental High School 1. Thus, for Experimental High School 1, the ARK program was found to be almost entirely successful. Encouragingly, at and across both experimental schools, no statistically significant gender differences emerged for any of the three ARK constructs and 17 ARK subcontracts, indicating that the ARK program was equally effective for the men and women teachers. Further, none of the three ARK constructs and 17 ARK subcontracts were statistically significantly related to (a) the number of students in the class, (b) the number of boys in the class, and (c) the number of girls in the class. Thus, the effectiveness of the ARK program did not vary as a function of class size or student gender composition.

However, just as the MMTBIE has answered some questions regarding the effectiveness of the ARK intervention, it has also raised some questions, including the following:

- Why was the ARK intervention relatively more effective for Experimental High School 1 than for Experimental High School 2? Did these differences in favor of Experimental High School 1 (across all 3 ARK constructs and 9 ARK subconstructs) stem from aspects of the ARK program itself? Did these differences stem from differences in how the ARK program was implemented at both experimental schools? Did these differences stem from differences in characteristics of the teachers at both schools (e.g., teaching experience of teachers)? Did these differences stem from differences in characteristics of the students at both schools (e.g., socioeconomic status, ethnic composition)?

- At Experimental High School 1, compared to the control school, why was the ARK program most effective with regard to classroom organization? Why was this program more effective for classroom organization than for emotional support? Did this difference stem from aspects of the ARK program itself? For example, does the ARK program emphasize classroom organization more than emotional support, which might be somewhat surprising bearing in mind that the mission of the ARK organization is “to teach adults to give the unconditional love that builds high self-esteem and a positive self-concept in children” (Wilkerson, 2007, para 1). Did this difference stem from any deviation from the protocol in how the ARK program was implemented at Experimental High School 1? Did this difference stem from differences in characteristics of the teachers at Experimental High School 1? Did this difference stem from differences in characteristics of the students at both schools?

- At Experimental High School 1, compared to the control school, why was the ARK program relatively ineffective with regard to instructional support? Did this ineffectiveness stem from aspects of the ARK program itself (e.g., less emphasis on instructional support)? Did this ineffectiveness stem from any deviation from the protocol in how the ARK program was implemented at Experimental High School 1? Did this ineffectiveness stem from differences in characteristics of the teachers at Experimental High School 1? Did this ineffectiveness stem from differences in characteristics of the students at both schools?

- At Experimental High School 1, compared to the control school, why was the ARK program effective for the nonverbal interaction, negative climate, instructional learning format, regard for adolescent perspectives, behavior management, and student engagement subconstructs? Did these differences in favor of Experimental High School 1 stem from aspects of the ARK program itself? Did these differences stem from differences in how the ARK program was implemented at both experimental schools? Did these differences stem from differences in characteristics of the teachers at both schools? Did these differences stem from differences in characteristics of the students at both schools?
• At Experimental High School 1, compared to the control school, why was the ARK program most effective for the nonverbal interaction subconstruct? Did these differences in favor of Experimental High School 1 stem from aspects of the ARK program itself? Did these differences stem from differences in how the ARK program was implemented at both experimental schools? Did these differences stem from differences in characteristics of the teachers at both schools? Did these differences stem from differences in characteristics of the students at both schools?

• At Experimental High School 1, compared to the control school, why was the ARK program ineffective for analysis and inquiry and instructional dialogue subconstructs? Did these differences in favor of Experimental High School 1 stem from aspects of the ARK program itself? Did these differences stem from differences in how the ARK program was implemented at both experimental schools? Did these differences stem from differences in characteristics of the teachers at both schools? Did these differences stem from differences in characteristics of the students at both schools?

During our meta-evaluation of the process and product of the MMTBIE (i.e., Phase 8), we came to the realization that many of these questions could have been answered if we had had the opportunity to have observed the ARK training. Although we had read the ARK manual (Wilkerson, 1995/2013, 2007) several times very carefully, as part of developing the CLASS-ARK-S, without observing the training directly, we do not know the extent to which all 10 lessons outlined in the ARK manual were covered during the ARK training, as well as the extent to which all the discussion ideas and classroom activities designed for each lesson were actualized. Armed with such knowledge, we would have been able to compare what we observed in the classroom to what was emphasized in the ARK training, and, as such, assess what Onwuegbuzie (2003) referred to as the level of implementation bias, (see also Benge, Onwuegbuzie, & Robbins, 2012). According to Onwuegbuzie (2003),

[[Implementation bias is a common and serious threat to internal validity in many educational intervention studies. Indeed, it is possible that implementation bias is one of the most frequent and pervasive threats to internal validity at the data collection stage in intervention studies. Implementation bias often stems from differential selection of teachers who apply the innovation to the intervention groups. In particular, as the number of instructors involved in an instructional innovation increases, so does the likelihood that at least some of the teachers will not implement the initiative to its fullest extent (Rogers, 1995). Such lack of adherence to protocol on the part of some teachers might stem from lack of motivation, time, training, or resources; inadequate knowledge or ability; poor self-efficacy; implementation anxiety; stubbornness; or poor attitudes. Whatever the source, implementation bias leads to the protocol designed for the intervention not being followed in the intended manner (i.e., protocol bias). For example, poor attitudes of some of the teachers toward an innovation may lead to the intervention protocol being violated, which then transgresses to their students, resulting in effect sizes being attenuated. A particularly common component of the implementation threat that prevails is related to time. Many studies involve the assessment of an innovation after one year or even less, which often is an insufficient time frame to observe positive gains (Rogers). Differences in teaching experience between teachers participating in the intervention and non-intervention groups is another way in which implementation bias may pose a threat to internal validity. (p. 77)]

Thus, we recommend that in future evaluations of the ARK intervention, the evaluators are given the opportunity to observe and to evaluate the complete ARK training, which would also include interviewing trainees as a means of assessing whether any of them suffer from "lack of motivation, time, training, or resources; inadequate knowledge or ability; poor self-efficacy; implementation anxiety; stubbornness; or poor attitudes" (Onwuegbuzie, 2003, p. 17) to implement the ARK tenets to the maximum effect. The ARK program trainers also could be interviewed to assess their perceptions of their trainees. Further, we recommend that the evaluators use both qualitative research techniques (e.g., observations, interviews, focus groups) and quantitative research techniques (e.g., developing an instrument such as a checklist to facilitate observations; developing a 4-point or 5-point Likert-format scale measuring each classroom teacher's impression of the strategies in the areas of effectiveness, efficiency, ease of implementation, application, and generalization) to assess the level of implementation of the ARK tenets by each teacher (cf. Collins, Onwuegbuzie, & Sutton, 2006). By so doing, the evaluators will be in a position to assess the level of implementation bias—that is, the discrepancy between the ARK program and the way it is implemented in the classroom. Then, during the data analysis process, the evaluators could correlate the implementation bias data with the CLASS-ARK-S data, so that findings regarding the effectiveness of the ARK program—vis-a-vis the three ARK constructs and 13 ARK subconstructs—can be placed in a more appropriate, generalizable, and transferable context. Additionally, to address Onwuegbuzie’s (2003) concern regarding time being a threat to identifying the effectiveness of the ARK program, we recommend that future evaluators conduct a follow-up evaluation of the teachers of the present sample members by assessing their teaching effectiveness over time.
Applying Change Theory to Future Implementation and Evaluation of the ARK Program

Our meta-evaluation of the process and product of the MMTBIE also has led to us to frame the future implementation and evaluation of the ARK program within the context of a leading change theory—namely, that of Rogers (2003). According to Rogers’s (2003) theory, diffusion is the process by which an innovation is transmitted, or diffused, via specific modes over time among the members of a socio-cultural system. Therefore, diffusion is a form of social change, that is, the process by which “alteration occurs in the structure and function of the social system” (Rogers, 2003, p. 6). This theory can be mapped onto the ARK program, wherein it serves as the innovation and the members of the socio-cultural system are represented by the teachers.

Rogers (2003) theorized that time is an important element of the diffusion process, and this is no different for the ARK program. Accordingly, time is represented by (a) the innovation-decision process by which the teacher passes from knowledge of the ARK tenets to its full adoption/embrace or rejection; (b) the relative earliness/lateness with which the ARK program is fully adopted/embraced by a teacher relative to the other teachers within the school; and (c) the rate of full adoption/acceptance of the ARK program, as measured by the number of teachers in the school who fully adopt/embrace the ARK program in a given time period. In the context of the ARK program, the innovation-decision process is the mental process through which a teacher passes from knowledge of the ARK program to forming an attitude toward it, to a decision to adopt fully or to reject one or more elements of the ARK program, and to confirmation of this decision. As such, there are five steps in the ARK program diffusion process: (a) knowledge, (b) persuasion, (c) decision, (d) implementation, and (e) confirmation. The evaluators then should obtain information from every teacher at various stages of the innovation-decision process in order to minimize uncertainty about the teacher’s expected outcomes.

According to Rogers (2003), there are five adopter categories that represent the innovativeness or a propensity to embrace fully all components of the ARK program. First are the innovators, who are the first teachers to embrace fully every component of the ARK program. Second are the early adopters, who are second fastest group of teachers who embrace fully every component of the ARK program, and who have the highest degree of opinion leadership among the other adopter categories inasmuch as they are able to influence informally other teachers’ attitudes toward the ARK program. Third is the early majority, who embrace fully every component of the ARK program after a varying degree of time. Fourth is the late majority, who embrace fully every component of the ARK program after the majority of the teachers do so, and they approach the ARK program with a high degree of skepticism about one or more elements of the intervention, and who show little opinion leadership. Last are the laggards, who are the last teachers in the school to embrace fully every component of the ARK program, if at all, and often who show little or no opinion leadership.

Based on the qualitative findings, it appears that the teachers with the least teaching experience—especially those in their first two years—tend to represent the innovators, whereas those beyond two years, but who are still early in their teaching careers and yet confident and secure enough to promote the ARK program tend to represent the early adopters. Those in the mid-careers tend to represent either the early majority or late majority. Finally, those teachers who are towards the end of the careers represent the laggards—as evidenced by the chemistry teacher at one of the experimental schools who was in her last year of teaching and who was one of the most critical of the ARK program. Thus, in the future, we recommend that the administrators of the ARK program initially target the innovators—that is, those teachers who are in the beginning of their teaching careers and who are the most impressionable. Then, after these teachers have been evaluated and modifications to the program have been made, the early adopters can be targeted for the ARK intervention, and so on, until there is complete buy-in by the vast majority of teachers in the school.

Further, based on the qualitative findings, we recommend that the ARK trainers make clearer not only the rationale and significance of the ARK program, but, even more importantly, also identify the uniqueness of this program—especially to teachers who represent the early majority and late majority. Additionally, we recommend that the ARK program designers re-evaluate both the ARK teacher covenants—especially the self-affirmations—and the ARK videos with a view to revising both elements. In so doing, it is essential that they field-test all revisions made.

In conclusion, the MMTBIE has provided compelling evidence that the ARK intervention thus far has been mostly successful in increasing the effectiveness of the select teachers at the two experimental schools, especially at Experimental High School 1. As such, the ARK intervention has much potential to effect positive change in high school classrooms. However, before this ARK intervention is delivered on a large scale, additional revisions and evaluations are needed using the evidence-based guidelines outlined in this report.
References


Appendix A: Classroom Assessment Scoring System—Adults Relating to Kids for Teachers and Staff Secondary Level (CLASS-ARK-S; Onwuegbuzie & Benge, 2014)

ARK Protocol
Teacher: _________________________________ School: ________________
Observer: ______________________________ Date: _________________

<table>
<thead>
<tr>
<th>Domain</th>
<th>Indicators</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Climate</td>
<td>• Student/Teacher relationships • Positive affect • Positive student/teacher communications • Respect • Use of humor • Knows all the students’ names • Uses appropriate physical touch with multiple students (e.g., High-five) • Maintains regular eye contact with all students as he/she talks to them • Provides focused (i.e., undivided) attention when interacting with students • Creates a nurturing and caring environment • Exhibits a passion for teaching and learning • Provides students the opportunity mentally to picture a positive future • Demonstrates being an advocate for students</td>
<td></td>
</tr>
</tbody>
</table>

Teacher Sensitivity
• Awareness • Responsiveness to academic and social/emotional needs and cues • Effectiveness to addressing problems • Separates the student from the behavior • Enables students to record good deeds undertaken by other students • Showing empathy for students (e.g., “I can understand why you feel this way …”) • Validating students’ feelings (e.g., “I want you to know that I really care about how you feel …”) • Offers the student choices • Provides power-seeking students with the opportunity to “save face”

(Continued)
<table>
<thead>
<tr>
<th>Domain</th>
<th>Indicators</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regard for Adolescent Perspectives</strong></td>
<td>• Flexibility and adolescent focus</td>
<td>• Atmosphere is created whereby students can feel wanted and accepted</td>
</tr>
<tr>
<td></td>
<td>• Connections to current students' lives</td>
<td>• Asking students to share their feelings</td>
</tr>
<tr>
<td></td>
<td>• Support for autonomy and leadership</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Meaningful student/teacher interactions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Atmosphere is created whereby students can feel wanted and accepted</td>
<td></td>
</tr>
<tr>
<td><strong>Nonverbal Interaction</strong></td>
<td>• Physical contact</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Eye contact</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Focused attention</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Active listening</td>
<td></td>
</tr>
<tr>
<td><strong>Behavior Management</strong></td>
<td>• Clear expectations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Proactive</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Effective redirection of misbehavior</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Applies age- and school-appropriate consequences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Not drawn into conflicts</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Reconnects with the student after discipline</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Avoids being defensive when confronted by misbehavior</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Connects the misbehavior (i.e., Attention-Seeking, Rebellious, Power-Seeking) with the student’s unmet need (i.e., to be loved and to belong; to be autonomous and free; for life to be meaningful and fun)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Uses self-evaluation technique to help misbehaving students to think about their misbehaviors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Provides attention-seeking students with choices</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Maintains a dispassionate demeanor and speaks with a calm, normal-volume voice when reacting to a rebellious student</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Allows power-seeking students to engage in mutual problem solving toward a “win-win” solution</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Provides power-seeking students with the opportunity to “save face”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Displays the use of conflict-resolution strategies when dealing with power-seeking students</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Refrains for completing tasks for power-seeking students that they are not undertaking due to avoidance behaviors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Focuses on power-seeking students’ strengths and not their weaknesses</td>
<td></td>
</tr>
</tbody>
</table>
Productivity
- Maximizing learning time
- Routines
- Transitions
- Preparation
- Prepares lesson that is organized, clear, and, whenever possible, entertaining

Negative Climate
- Negative affect
- Punitive control
- Disrespect
- Humiliates
- Teacher exhibits lack of enthusiasm
- Teacher exhibits lack of focus
- Teacher exhibits frustration
- Teacher appears to be overwhelmed
- Negative reinforcement

Instructional Learning Formats
- Learning targets/organization
- Variety of modalities, strategies, and materials
- Active facilitation
- Effective engagement
- Exhibits high but realistic expectation for each student
- Emphasizes rigor

Content Understanding
- Communication of concepts and procedures
- Background knowledge and misconceptions
- Transmission of content knowledge and procedures

(Continued)
Table. Continued.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Indicators</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis and Inquiry</td>
<td>• Facilitation of higher-order thinking</td>
<td>• Provides students with sufficient time to answer questions</td>
</tr>
<tr>
<td></td>
<td>• Opportunities for novel application</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Metacognition</td>
<td></td>
</tr>
<tr>
<td>Quality of Feedback</td>
<td>• Scaffolding</td>
<td>• Encourages students to maximize the quality of their work</td>
</tr>
<tr>
<td></td>
<td>• Building on student responses</td>
<td>• Seeks feedback from students on an ongoing basis to gauge levels of understanding</td>
</tr>
<tr>
<td></td>
<td>• Encouragement and affirmation</td>
<td>• Seeks feedback from students at the end of the class</td>
</tr>
<tr>
<td></td>
<td>• Encourages students to maximize the quality of their work</td>
<td>• Administer a brief evaluation sheet for students to use to evaluate the class</td>
</tr>
<tr>
<td></td>
<td>• Seeks feedback from students on an ongoing basis to gauge levels of understanding</td>
<td>• Makes it clear that feedback from students is taken very seriously</td>
</tr>
<tr>
<td>Instructional Dialogue</td>
<td>• Cumulative content-driven exchanges</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Distributed talk</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Facilitation strategies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Uses story-telling techniques</td>
<td></td>
</tr>
<tr>
<td>Student Engagement</td>
<td>• Active engagement</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Monitors the classroom to ensure that all students are engaged</td>
<td></td>
</tr>
</tbody>
</table>

Note: Adapted from Secondary CLASS observation tool and ARK (Adults Relating to Kids) for Teachers: Secondary Level (Volume 1) by Dr. B. G. Wilkerson.