Development and Testing of Items Measuring School Staffs' Perceptions of Aligned and Balanced Curriculum

Merrill L. Meehan, AEL
William Wiersma, University of Toledo
Kimberly S. Cowley, AEL

AEL
P. O. Box 1348
Charleston, WV 25325

Paper Presented at the 13th Annual CREATE National Evaluation Institute
Colorado Springs, Colorado
July 8-10, 2004

Founded in 1966 as a not-for-profit corporation, AEL provides services to educators, education publishers, and policymakers. Services include rigorous research design and implementation, research reviews, intensive product and program evaluations and randomized field trials, technical assistance, and award-winning professional development programs. AEL operates several contracts funded by the U.S. Department of Education: a Regional Educational Laboratory, the Region IV Comprehensive Center, and an Eisenhower Regional Consortium for Mathematics and Science Education.

Information about AEL projects, programs, and services is available by contacting AEL.

Post Office Box 1348
Charleston, West Virginia 25325-1348
304-347-0400
800-624-9120
304-347-0487 (fax)
aelinfo@ael.org
www.ael.org

This publication is based on work sponsored wholly or in part by the Institute of Education Sciences (IES), U.S. Department of Education, under contract number ED-01-CO-0016. Its contents do not necessarily reflect the views of IES, the Department, or any other agency of the U.S. government.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Increasingly, academics and practitioners are looking at nonstructural aspects of schooling as the "doors" to educational improvement (Joyce, 1991). Such doors to school improvement include the shared norms, knowledge, and skills of teachers (Elmore, 1995). Restructuring of schools, therefore, needs to be balanced by "reculturing" of school faculties (Fullan, 1996; Hargreaves, 1994). Thus, while many present reform efforts concentrate on school results as evidenced by students' test scores, there is a need to be able to measure and report on the school staff's perceptions of their abilities to move into and maintain a mode of continuous learning and improvement, which is one form of reculturing.

AEL, in its role as a regional educational laboratory, has been committed to research on school improvement since 1966. Among AEL projects was Quest (1996-2000), a network of school communities located in Kentucky, Tennessee, Virginia, and West Virginia. Quest schools were dedicated to building and sustaining learning communities that supported high levels of student ***and*** adult performance. The Quest network of schools emphasized six key components or dimensions in its conceptual framework. Those six key components were: shared leadership, effective teaching, school/family/community connections, purposeful student assessment, shared goals for learning, and learning culture.

The *AEL Continuous School Improvement Questionnaire*

In an effort to assist school staff toward improvement, especially continuous learning and improvement, AEL needed an instrument to measure school staff perceptions of their status. AEL's conceptual framework of six key components became the basis for the development of the needed instrument. The resultant instrument would not only inform a school staff of the extent to which they perceived themselves as a high-performing learning community but, when compared with scores of staff in schools know to be institutions fostering high performance levels of both students and adults in the school, the results could inform them of how accurate those perceptions are based on the normed data.

Starting in spring 2000, the AEL research and evaluation staff began the development, pilot testing, and field testing of the *AEL Continuous School Improvement Questionnaire* (AEL CSIQ) (Meehan, Cowley, Wiersma, Orletsky, Sattes, & Walsh, 2002). For the pilot test of the AEL CSIQ, 147 items unevenly distributed across the six dimensions in the conceptual framework were tested in 28 schools in the four AEL states. The main purpose of the pilot test was to reduce the length of the total instrument by selecting an equal number of the best items for each of six dimensions. Based on the descriptive statistics from 274 professional staff in the pilot test, the AEL CSIQ was reduced to 12 items per subscale for a total of 72 items. To assess the concurrent validity of the AEL instrument, a similar instrument also was administered.

The first field test of the 72-item version of the AEL CSIQ was conducted in the fall and winter of 2000. The 6-point response option scale of "Is not present" to "Is present to a high degree" remained the same as in the pilot test. A total of 2,093 mostly teacher respondents in 79 schools in AEL's four states completed and returned the instrument. The schools volunteered to complete the AEL CSIQ in exchange for an individual school report of their results or they were

volunteered to complete the instrument by others because they were on notice to make improvements in their students' performance levels on the statewide testing program.

Then, in an effort to make the AEL CSIQ more convenient for respondents but still retain satisfactory internal consistency reliabilities for the six subscales and the total score, the instrument was shortened once again by dropping two items per subscale. Also, the remaining 60 items were placed in random order. This 60-item version of the AEL CSIQ was administered to the full faculties of 75 schools, of various levels, in Tennessee. These schools were participating in a school improvement project with AEL staff and they also received individual school reports of their results. Also, a subset of schools in the second field test volunteered to complete a second copy of the instrument two or three weeks after their first completion for the purpose of determining the test-retest (stability) reliabilities.

In summary, through multiple tests, the AEL CSIQ was reduced from 147 items with unequal numbers of items in subscales to a more manageable and convenient length of 10 items per each subscale. The AEL CSIQ was found to be very sound technically. Internal consistency reliability was high for all scales, as was the stability reliability. The instrument showed satisfactory concurrent validity with another school climate instrument. The items possessed face validity and the factor analyses provided strong support for the construct validity of the entire inventory as reflected by the subscales (Meehan, Cowley, Wiersma, Orletsky, Sattes, & Walsh, 2002, p.22).

Norming the AEL CSIQ

Based on satisfactory reliability and validity results from the series of pilot and field tests, the AEL CSIQ was put into practice in several school improvement projects. In a short time, 132 schools had administered the instrument to their professional staff. An emerging need was for norms for the instrument to make the interpretation of its scores more meaningful. In 2002, AEL research and evaluation staff completed a norming study of the AEL CSIQ based on the number of schools who had completed it at that date (Meehan, Wiersma, Cowley, Craig, Orletsky, & Childers, 2002).

The purpose of the study was to report normative AEL CSIQ data for the total of 132 schools who had completed it by 2002. The normative data were developed and reported by type (level) of school, locale type (Johnson) codes, and schools nominated to be high performing learning communities. This latter group of schools requires description. Within the total of 132 schools, there was a special subgroup of schools who were nominated by either AEL or Tennessee Department of Education staff as being high-performing schools *and* professional learning communities. These special schools, which are referred to as the "Known" schools, were viewed as possessing positive characteristics relative to continuous learning and improvement by <u>both</u> the students in the school and by the adult professional staff. These special schools were nominated to possess the characteristics of high-performing learning communities—there was no guarantee that they were, in fact, high-performing learning communities. Indeed, one of the chief purposes of the norming study for the AEL CSIQ was to study the normative data from the set of "Known" schools (Meehan, Wiersma, Cowley, Craig, Orletsky, & Childers, 2002, p. 7).

The norming of the AEL CSIQ resulted in several conclusions. One, the type of school, that is level (elementary, middle, high), appears to have a slight to modest effect on the AEL CSIQ subscale and total score performance. Respondents in elementary schools and schools with elementary grades (PreK-12) had high scores. Two, there is no evidence that scores on the AEL CSIQ related to the extent of rurality-urbanicity of the school locale. Third, educators in schools nominated to be high-performing learning communities on the basis of their commitment to continuous learning and improvement of both students and staff almost always score higher on the AEL CSIQ subscales and total score than their counterparts in the remaining schools of the same type. The single exception at the middle school level was explainable by the fact there only was one "Known" middle school. Fourth, the patterns of scores showing the respondents in the nominated "Known" schools having greater AEL CSIQ subscales and total score means (except the middle school group, as explained) than of respondents in the remaining schools supports the AEL assumption that a faculty's commitment to continuous learning and improvement is the critical dimension in defining schools as high-performing learning communities (Meehan, Wiersma, Cowley, Craig, Orletsky, & Childers, 2002, pp. 32-33).

Research Involving the *AEL CSIQ*

Achievement gap research. Since the norming of the AEL CSIQ with 132 schools, it has been administered to approximately 500 other schools as a part of various research, evaluation, and school improvement projects. Several of these projects are ongoing multiple-year efforts for which the AEL CSIQ results are used as formative information for project directors and school staff to employ in making decisions about directions or courses of action to take to assist in bringing about school improvement. Other projects employing the AEL CSIQ are completed and are discussed briefly below.

A study by Cowley and Meehan (2002) investigated the differences among professional staff's commitment to continuous learning and improvement in high-performing schools that were differentiated by student academic performance disaggregated by race and socioeconomic status. The objectives of this study were to (1) study and compare the descriptive statistics on the AEL CSIQ for those schools with minimal achievement gap differences by subgroup and those with large achievement gap differences by subgroup, (2) determine whether significant differences existed between those minimum and large gap schools, (3) compare the descriptive statistics for schools by building level, and (4) determine if significant differences existed among building levels.

The 48 high-performing schools in the study were identified by staff at the Kentucky Department of Education from the population of all public schools in the state (approximately 1,400). All 48 schools were identified as being relatively high performing based on their overall academic school index scores. One half of them also were relatively successful with struggling learners and minority and economically disadvantaged students. The other half of the schools were not as successful with the identified subgroups. The 24 schools per group included 12 elementary, 6 middle, and 6 high schools (Cowley & Meehan, 2002, p.4). Data were received from 47 of the 48 schools.

Overall, across the two groups of schools (minimum gap and large gap) the AEL CSIQ school/family/community connections scale received the lowest ratings for each building level.

Conversely, the effective teaching scale received the highest mean ratings at elementary and high schools, also with the two lowest standard deviations. Looking within achievement grouping, the large-gap middle and high schools and the minimum-gap middle schools showed more cohesion and less dispersion in their perceptions than their counterparts. The GLM ANOVA also revealed significant main effects by achievement group for three of the six scales. Schools with minimum achievement gap differences had significantly higher scores for learning culture, shared goals for learning, and effective teaching than those with large differences in achievement gap (Cowley & Meehan, 2002, p. 13). Shared leadership was the only scale in which statistical significance was not detected by either building level or achievement gap. Further, no significant interactions were found between building level and achievement gap.

Findings of this study suggest wide variations in professional staff's commitment to continuous learning and improvement between achievement gap groups and across building levels. Results suggest that the area of school/family/community connections is one area that may be most in need of intervention for schools in general if they aspire to become high-performing learning communities. "If schools are trying to make yearly progress toward meeting the needs of all students, it is not enough to focus on structural changes, new standards, or accountability requirements," Cowley and Meehan write (2002, p. 15). Their study suggests that attention also must be given to fostering and sustaining a school climate where the professional staff are committed to continual learning and improvement. For schools studied with achievement gaps, this is especially true in the areas of learning culture, shared goals for learning, and effective teaching.

Cross-state study. A second study by Meehan and Cowley (2003) employed the AEL CSIQ to investigate low-performing schools, high-performing schools, and high-performing schools in two states. Two samples of schools were used in this study, both identified by staff in their state departments of education (SDE). Both states were in the southern region of the United States. A sample of 45 schools, out of a population of 1,470, were identified by their SDE as low-performing based on their SAT-9 scores and a sample of 47 high-performing schools were identified by their SDE based on their overall academic score on the state testing system. The AEL CSIQ was administered to the professional staff in the winter of 2002.

Four objectives guided the data analyses and results of this basically descriptive study. The first objective was to identify those schools within each state with scores consistently above or below the median on the six scales of the AEL CSIQ. The second objective was to inspect the range of scores within each state for any overlap between the two groups of consistency above and below the median for each scale. The third objective was to identify from the high-performing schools those that are classified as high-performing learning communities. The fourth objective was to study and compare the descriptive statistics of the scale scores across the states (Meehan & Cowley, 2003, p. 2).

Conclusions from the cross-state study were interesting. Within a sample of schools identified as being low performing from one state, the AEL CSIQ differentiated between the school professional staff's level of commitment to continuous learning and improvement on all six scales. Similarly, within a sample of school identified as being high performing from another state, the AEL CSIQ differentiated between the professional staff's level of commitment to continuous learning and improvement on the six scales. Across the states, professional staff identified the areas of school/family/community connections as being the area most in need of learning and improvement compared to the other five areas measured by the instrument. Even though the mean scores on the AEL CSIQ tend to be rather high on the 60-point scale (which may be a function of the self-report nature of the instrument) and despite the rather narrow spread of scores across scales and samples of schools, nonetheless, the instrument does differentiate professional staff's commitment to continuous learning and improvement within schools similarly classified in terms of their academic performance (Meehan & Cowley, 2003, p. 16).

Professional staff in schools identified as being high performing on the basis of students' academic performance always scored higher on the AEL CSIQ scales than the professional staff in schools identified as being low performing on the basis of students' academic performance. Therefore, Meehan and Cowley concluded that measuring a faculty's commitment to such continuous learning and improvement is one effective way to assess the reculturing of the school's professional staff (2003, p. 16). Finally, from the cross-state study, "assuming the key components of high-performing learning communities to be high levels of student achievement ***and*** professional staff's commitment to continuous learning and improvement, this study showed that high-performing schools are not necessarily high-performing learning communities" (p. 17).

Purpose and Objectives

The AEL CSIQ instrument has been administered to many schools in its short history, both for research and school improvement efforts. The instrument has proven useful in describing high-performing learning communities. However, given the current emphasis by the states to both develop state curriculum standards for students in many key content areas and then the emphasis of developing state testing programs keyed to their state standards, it became apparent that the AEL CSIQ does not include the area of curriculum within the instrument's conceptual framework. Thus, while many states emphasize—even require—that some school improvement efforts, such as technical assistance and professional development programs, include the component of state curriculum standards, the AEL CSIQ did not address this area directly. AEL staff saw the need to measure the K-12 school staff's perceptions of aligned and balanced curriculum as part of the AEL CSIQ. The main purpose of this study, then, was to develop a seventh scale to the AEL CSIQ on the topic of aligned and balanced curriculum.

Four objectives guided this study. The first objective was to develop a set of draft instrument items based on a framework for aligning and balancing the school curriculum. The second objective was to review, refine, and reduce the draft items into a pilot-test instrument. The third objective was to administer the pilot test instrument to K-12 school faculties of different levels (elementary, middle, high) and types (high performing or not). Fourth, through statistical analyses, the objective was to reduce the instrument to 10 items for the field test.

METHODS

This section describes the methods of the study, including the sample, the instrument, the data collection, and the data analyses.

Sample

Schools. Full K-12 school faculties were recruited to participate in the pilot test of the aligned and balanced curriculum scale items. In exchange for participating in the pilot test, the principals or project directors were offered individual school reports of the six established AEL CSIQ scales as an incentive. Thus, the sample of schools in the pilot test were volunteers or were volunteered by someone directing a school improvement project, in which the school was participating.

A total of 86 schools in California, Kentucky, Tennessee, Virginia, and West Virginia agreed to participate in the pilot test. By school level, there were 50 elementary, 21 middle, 4 high, 7 K-8, 2 K-2, and 2 vocational schools in the study. As part of the design of the pilot test, a portion of the schools were recruited specifically to offset the majority of the schools that were identified as being low performing schools, at one time, because they were in a statewide school improvement project. These 13 special schools were identified as being high-performing within their respective states. Typical criteria to be labeled high performing within a state included such things as being named a "School of Excellence" by that state, being named a "Blue Ribbon" school in that state, or some similar such designations.

Respondents. A total of 2,459 returned pilot test instruments were utilized in this study. The vast majority of these respondents (n = 2,126) were located in the large set of schools that were not identified as being high performing (labeled as "Other"). The remaining respondents (n = 333) were professional staff in schools identified as being high performing on some recognized, statewide basis. Although the split across the two groups of schools provided uneven frequencies, 86.5 percent in the former group and 13.5 percent in the latter group, this was not a problem in the study because the frequencies were large. As stated above, whole school professional staff were asked to complete the pilot-test instrument as well as the regular, 60-item AEL CSIQ.

Respondents in both groups of schools were asked to complete a set of demographic questions as part of the field test. Not all respondents answered all the demographic items, so the valid responses vary by item. However, the number of skipped items was rather small across the two groups, ranging from 5 to 12 for the high-performing group and from 35 to 82 for the Other group.

Regarding the respondents in the Other schools group, 73.5% were regular classroom teachers, 7.1% were special education teachers, 3.6% were principals/assistant principals, 2.7% were counselors, 2.0% were librarians/media specialists, and 11.1% checked "other" for role. In terms of formal education, a little more than half of the group had either a bachelor's degree (27.9%) or a master's degree (26.8%), 15.1% had a master's degree plus 30 or more credits, and 11.4% had a bachelors degree plus 15 credits, and 9.3% had a bachelor's degree plus 30 or more

credits. Over three fourths (83.7%) were females and 16.3% were males. Black or African-American was checked by 51.3%, White by 43.9%, "Other" by 2.7%, and the remainder across the other ethnic groups in the list. Nearly all of the respondents in the Other school group, 98.8%, worked in elementary schools. Respondents in the group were rather experienced educators. They reported working an average of 14.6 years (SD = 10.0) in any school, an average of 10.2 years (SD = 9.0) in their present school district, and an average of 8.4 years (SD = 6.7) in their present school.

With respect to the respondents in the high-performing schools groups, 72.5% were regular classroom teachers, 13.0% were special education teachers, 2.8% were principals/assistant principals, 1.9% were counselors, 1.2% were librarians/media specialists, and 8.6% checked "Other" for role. In terms of formal education, 29.5% had a master's degree plus 30 or more credits, 22.0% had a master's degree, 16.8% had a bachelor's degree, 6.8% a bachelor's degree plus 30 or more credits, 6.5% a master's plus 15 credits, 0.3% a doctorate, and 2.8% marked "Other." Just over two thirds (67.3%) of the respondents in this group were females and 32.7% were males. Nearly all the respondents (97.8%) were White, 1.2% were Black or African American, and 0.3% each marked three other ethnicity categories. As expected, the respondents were nearly evenly distributed across school levels with 28.0% elementary, 33.5% middle, and 34.8% high school. Respondents in the high-performing school group were more experienced than those in the Other school group. Educators in this group had worked an average of 16.9 years (SD = 10.4) in any school, an average of 15.2 years (SD = 10.6) in their present district, and an average of 11.8 years (SD = 9.5) in their present school.

Instrument

The development of the pilot-test version of the aligned and balanced curriculum instrument followed a traditional path. First, a published framework for aligning and balancing the curriculum (Ceperley & Squires, 2000) was provided to a pair of very experienced curriculum consultants to be used as the organizing structure for the drafting of individual items. The pair of curriculum developers each prepared a set of draft items, which numbered 47 items at this stage. These 47 initial items were reviewed by a technical advisor to the project who eliminated redundancies and improved technical deficiencies, such as dual or triple concepts within a single stem. Draft items were reviewed, deleted, or revised so that 35 items remained. These remaining 35 items were at least technically adequate for the pilot test. The response options for all the items was a 6-point, Likert-type scale from 1 = "Not present" to 6 = "Present to a high degree," just like the response options for the other 60 items in the AEL CSIQ.

The 35 pilot-test items were assembled into a single page (printed front to back) supplement to the larger AEL CSIQ instrument. Both the supplement and the AEL CSIQ instrument were printed to be scanned by AEL's equipment. Because the supplement was loose and the "regular" instrument that it was inserted into was a fold-over instrument, unique paper label bar codes were affixed to each, in case they became separated during the handling.

Data Collection

The pilot test of the aligned and balanced curriculum items, as a supplement page to the "regular" AEL CSIQ instrument, took place during the 2002-2003 school year. Not all schools participated in the pilot test during the same months within the academic year, but all data collection ceased by May 2003. Procedurally, packages of the assembled instrument were shipped to a school contact person for distribution, collection, and shipping back to AEL for analyses. Each package of instruments included a return, addressed envelope to AEL.

Returned packages of completed instruments were tracked by AEL staff utilizing a spreadsheet. Packages were opened, instruments counted, checked, and logged into the database. Next, the individual school's instruments were scanned into the Remark scanning program, then exported to SPSS for analyses. Still at the individual school level, the instruments' data were analyzed in SPSS, tables constructed in Excel software, and the results were imported into a Word template profile for each school. AEL staff analyzed the school's data in the AEL CSIQ profile for the six "regular" scales only and typed the interpretation of the data into the school report. These individual school reports were then shipped back to the school contact person for dissemination and utilization at the school level. This was completed for all 86 schools in the pilot test.

Data Analyses

The first step in the data analyses was to merge all the 86 individual school files into one aggregated file in SPSS. Next, descriptive statistics on the 35 items were computed. Then, internal consistency reliability estimates were generated. Factor analysis of the 35 items by the full group of respondents was computed using principal axis with Varimax and Kaiser normalization rotation. In this step, exploratory factor analysis was employed, not confirming the number of factors. This was followed by computing correlations of the 35 items to the other six established scales of the AEL CSIQ. Then, the full file was split by the respondents in high-performing schools and those in the Other schools and new descriptive statistics, reliability estimates, and factor analyses were computed. Last, comparisons of item means across the two groups of schools were made.

FINDINGS AND DISCUSSION

The purpose of the pilot test of the 35 items on aligned and balanced curriculum was to produce a scale of ten items to be used as the basis for the field test of the scale. The goal of the pilot test was to find the ten "best" items that would make up a scale to be added to the six established scales in the AEL CSIQ instrument.

The purpose of this section is to present the results of the pilot test of the 35 aligned and balanced curriculum items. To accomplish this, the results are presented in two main headings: by the respondents in the full group of 86 schools in the pilot test, and then by those schools broken into the two groups of high-performing schools and Other schools. Also, the presentation of the results in this section is accompanied by a discussion of those results. Each major heading below is organized by the various statistical techniques employed with the data.

Full Group of Schools

Descriptive statistics. Inspection of the descriptive statistics from the full group of schools shows some of the 35 pilot test items essentially had no discrimination value, even though they appeared to have content validity. Item numbers 1 and 2 serve as examples of this with items means of 5.42 and 5.37 respectively on the 6-point scale and standard deviations around .90. Both of these items had medians of 6.00, indicating that at least half of the respondents gave the maximum rating of 6.

If AEL staff were to select the 10 items with the lowest item means, they would include 5, 6, 15, 16, 20, 24, 26, 33, 34, and 35. These items had means ranging from 3.48 to 4.42. Only two of these means were less than 4.0. Item numbers 6, 30, and 31 had means of 4.42, so any one of these could be included as the tenth item. These 10 items had standard deviations ranging from 1.1 to 1.6 with 9 being 1.3 or greater. The total mean score for these 10 items would be 41.48, a mean about 2.5 points less than the total score mean of the 10 items based on the factor analysis results (see below). Five of the 10 items with the lowest means were included with the 10 items based on the factor analysis results.

Actually, excluding items with means greater than 5.0 and medians of 6.0, there probably would be little difference in the technical characteristics of any combination of 10 items. Reliability estimates would be quite high, certainly acceptable (see below). The content of the items could be considered, but remaining items based on a content analysis would be a subjective procedure.

Reliability estimates. Even with 35 items, the new curriculum subscale showed great homogeneity. The internal consistency reliability estimate (Cronbach Alpha) for all 35 items was .97. This indicates that, practically speaking, selecting any 10 items would result in a subscale with adequate internal consistency reliability, most likely between .90 and .94.

One approach to selecting items to be retained when reducing the number of items is to select those with the greatest item-to-total score correlations. This approach would give the

greatest reliability for the remaining items. Employing this approach would retain items 18, 19, 22, 25, 26, 27, 28, 31, 33, and 34.

Factor analysis. A factor analysis was computed for the 35 pilot test items. A principal axis factor analysis using Varimax rotation with Kaiser normalization yielded 5 factors with eigenvalues greater than 1.0. Together, the 5 factors accounted for 60% of the total variance. Seventeen items had pattern/structure coefficients greater than .30 with Factor 1. Selecting the 10 items with the greatest pattern/structure coefficients retains items 23, 24, 26, 27, 30, 31, 32, 33, 34, and 35. Five of these items overlapped with items in the group of 10 items having the greatest item-to-total score correlations. For this second set of 10 items whose selection was based on the factor analysis, the item-to-total score correlations ranged from .68 to .77. The internal consistency reliability estimate for these 10 items was .93.

Considering the 10 items retained based on the factor analysis, the smallest mean was 3.97 and the largest was 4.97, exactly one point higher. All the item standard deviations were greater than 1.0, ranging from 1.29 to 1.48. Seven of these items had medians of 5.0 and three had medians of 4.0. The mean of the total score was 43.99, and the total score had a standard deviation of 10.60. One of the undesirable characteristics of the AEL CSIQ is that even staff in low-performing schools tend to score towards the upper ends of the measurement scales. For the total scores, the maximum possible score on a subscale is 60. The desirable situation would be to have means around the middle of the scales for both individual items and for the total score. Since the lowest score on an item is 1.0, the middle score is around 3.5 for the subscales, and the middle of the total score scale is around 35. Substantial variability also is desirable, especially among respondents from different types of schools.

If the aligned and balanced curriculum subscale is viewed as possessing multiple constructs, then the results of the factor analysis could be viewed another way to choose the 10 items for use in the field test. Recall, five factors had eigenvalues greater than 1.0 and we could select a pair of items from each factor with the largest pattern/structure coefficients. Doing this would yield items numbered 3, 4, 14, 15, 17, 18, 28, 29, 33, and 35. One exception was made in this hypothetical draw of items. Items numbered 1 and 2 had the largest loadings on Factor 4, but for reasons stated above in the descriptive statistic results, items 1 and 2 are not acceptable for inclusion in the shorter curriculum subscale.

With respect to the two items for each factor method to comprise the 10-item subscale, their item level means ranged from 4.09 to 4.95. Six of these items had means greater than 4.50. The subscale total score based on these 10 items would be around 45.5, a total score that is undesirably high. Two items had medians of 4.0; the remainder had medians of 5.0. All of their standard deviations were greater than 1.0, ranging from 1.08 to 1.47. The item-to-total score correlations ranged from .57 to .74. As a group, these correlations are lower than those of the items based on the first factor, but the correlations should be high enough to yield an adequate internal consistency reliability of a subscale made up of these 10 items.

Correlations. The correlations of the new items to the original AEL CSIQ subscales provides additional information. Here, the total scale score for the 35 aligned and balanced curriculum items was correlated to each of the original six subscales (see Meehan, Cowley, Craig, Balow, & Childers, 2002). These correlations ranged from .63 to .77 and were similar to the intercorrelations among the other scales. This shows that subscales of the AEL CSIQ are

correlated—even any combination of items selected for the newest subscale—and they are not independent from one another.

Discussion. All things considered from the data from the full group of schools, the selection of the 10 items that comprise the "best" set for the aligned and balanced curriculum subscale is somewhat arbitrary. Almost any selection of 10 items will result in adequate subscale internal consistency reliability, but also will tend to have an undesirably large total score mean. If we retain the 10 items based on the largest pattern/structure coefficients on Factor 1, the item with the greatest mean of those 10 items is number 23 with a mean of 4.98. If it were replaced with item 20 with a mean of 3.48, the total scale score would be lowered by 1.5 points. Item 20 had a large standard deviation, but it did not load on Factor 1.

Continuing the discussion, suppose one of the sets of 10 items is retained based on the results of the factor analysis of the full group data. Technical characteristics of the resulting subscale probably would be similar and would not provide a good basis for discriminating between the two sets. Due to the method of selection, there is an overlap of only two items between the sets, the last two items that have the largest pattern/structure coefficients on Factor 1. In the final analysis, the decision comes down to "What is most desirable, a scale heavily loaded on one factor, or one whose item loadings are distributed somewhat across multiple factors?' It may very well be that K-12 curriculum experts may have a preference related to the intended uses of the AEL CSIQ with respect to assisting in school improvement efforts/programs for a school. Put another way, with seven potential areas for initiating and fostering school improvement efforts results from administration of the AEL CSIQ, a school improvement specialist may choose one of the other six areas to start with and it may not be the curriculum area.

Two Educator Groups

Descriptive statistics. Total scores were computed for the 35 pilot test items for both groups. For the larger group of educators in Other schools, 1,702 respondents (80.1%) had total scores; for educators in high-performing schools 272 respondents (82%) had total scores. This shows that almost 20% of respondents in each group omitted at least one item out of the 35.

Individual item means were compared across the two groups of educators. The minimum and maximum scores on any item were 1 and 6. Inspection of the item mean differences across the two groups revealed 10 items that had differences greater than .60. Table 1 displays those items with mean differences greater than .60 between the respondents in the two groups. These 10 items included 5, 6, 8, 20, 23, 24, 26, 27, 31, and 34. The sum of these differences is 6.94 points. In comparing the group of 10 items with groups of 10 items identified earlier, there are the following overlaps:

- with greatest item-to-total score correlation, 4 items: 26, 27, 31, 34
- with greatest factor loadings on one factor, 6 items: 23, 24, 26, 27, 31, 34
- with lowest item means, 6 items: 5, 6, 20, 24, 26, 34

Thus, these 10 items in Table 1 that maximize the differences between the two groups have between 4 to 6 items in common with those items selected using other criteria.

Table 1: Items and Differences in Means Between the Two Groups for Items with a Difference Greater Than .60

Pilot Test Item Number	Difference in Means of Two Groups
5	.80
6	.82
8	.70
20	.65
23	.61
24	.62
26	.62
27	.79
31	.65
34	.68
	Sum = 6.94

The minimum and maximum scores on a new aligned and balanced subscale of 10 items are 10 and 60 points. If we estimate total score means for these two groups for the 10 items with greatest item differences, we estimate the mean for the high-performing schools would be 48.91 and that for the Other schools would be 42.00. These are estimates, computed by the sums of the means on the 10 items, and because of omissions, the numbers of respondents across items varied slightly. The 6.91 point difference between the scores corresponds with the sum of differences in Table 1.

Figure 1 is a graphic depiction of two subscales constructed by adding the items with greater mean differences across the two groups.

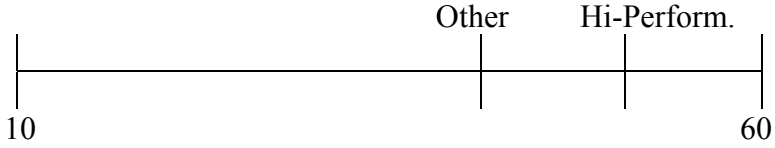

Figure 1: Graphic Depiction of Greatest Difference Subscale Means of the Two Groups on the Possible Subscale Score

The end points in the figure are the lowest possible score of 10 and the highest possible score of 60. This figure clearly shows how the means for both the Other schools group and the high-performing schools group were located near the high end of the possible score. At the same time, the figure also illustrates that there is separation of the two possible groups' scores with that of the high-performing group being higher.

Next, in descriptive statistics, the means for the total score on all 35 pilot test items were computed. The range of possible total scores was from 35 to 210 points. The total score mean for the high-performing schools group was 176.01 and for the Other schools group was 151.56. If we divide the difference in means of 18.45 points by 3.5, we obtain the value of 5.27 points. This value represents the average difference of the 10 items. The difference of 6.94 for the 10 items with the greatest differences in item means (see above) is substantially more than the average difference. The standard deviations for these total scores were 29.76 and 25.12, with the Other schools group being the most variable. Overall, though, these standard deviations indicate similar variance within the two groups.

Last, an analysis was computed for the difference between the means of the two groups on each of the 35 pilot test items. With a very high level of statistical power, all the differences were statistically significant, even beyond the .001 level. These were not random samples, but the t-test for the differences between the means show only that variance between the groups was much greater than variance within groups. Although the differences were modest, the high-performing schools group had the highest mean on all 35 items.

Reliability estimates. The internal consistency reliability estimates (alpha) for educators in both groups were very high. For the total scale score, both groups had Cronbach alphas of .97. Practically speaking, any ten items selected for the aligned and balanced curriculum scale would have the same adequate reliability, most likely in the .90s. So, internal consistency reliability is not really an issue for either group of educators in the pilot tests.

Factor analyses. Factor analysis was conducted on the scores from the two groups of respondents. The results were interesting. Considering the first factor extracted in the rotated factor matrix for each group and the 10 items with the highest pattern/structure coefficients (in descending order) on that factor, only one item appears in both sets of items—item number 23. The interpretation of this result is not entirely clear, but it seems that the major curriculum constructs for the two groups were somewhat different. Of course, the intercorrelations among the items were quite substantial, so even if the constructs are different, they are not very independent. However, this result does provide an argument against selecting the 10 items for the new aligned and balanced curriculum subscale—which is intended to be used across a wide variety of schools—on the basis of an overall factor analysis.

Correlations. The total scores for the 35 pilot-test items of the proposed aligned and balanced curriculum subscale were correlated with the scores on the other six scales of the AEL CSIQ. These new subscales were computed separately for the two groups in the pilot test. Table 2 displays these correlations. These correlations were moderately high and they also were consistent across the two groups. Too, they were consistent with the correlations among the other established six subscales (Meehan, Cowley, Craig, Balow, & Childers, 2002).

Table 2: Correlation Coefficient Between Total Score on the Curriculum Subscale and the Other Subscales Within Group

Subscale Name	Group	
	Other Schools	Hi-Performing Schools
Learning Culture	.76	.78
School/Family/Community Connections	.73	.67
Shared Leadership	.61	.54
Shared Goals for Learning	.76	.76
Purposeful Student Assessment	.76	.77
Effective Teaching	.73	.73

CONCLUSIONS AND RECOMMENDATIONS

The pilot test of a set of 35 items in the area of aligned and balanced curriculum was based on a large number of respondents in many schools. A variety of statistical analyses were computed on the respondents' scores in an effort to determine the 10 test items to comprise a subscale to add to the other six subscales in the AEL CSIQ instrument. Based on this pilot test of the 35 potential items, the following conclusions and recommendations are warranted.

1. Attaining adequate subscale internal consistency reliability should not be a problem, regardless of the 10 items selected for the final version of the aligned and balanced curriculum subscale.

2. Unless there is a compelling reason not evident in this pilot test, the 10 items retained to comprise the new subscale should be those showing the greatest difference between the means of the high-performing and the Other school groups. These items are listed in Table 1 above. These 10 items overlap some with those selected by other criteria, most importantly those with the lowest item means.

3. Much like the scores on the six established subscales of the AEL CSIQ, those of the new aligned and balanced curriculum subscale tend toward the high end of the measurement scale. This may limit the discrimination power of the new subscale but that power still may be adequate, at least consistent with that of the other subscales. Apparently, the AEL CSIQ items reflect extensive activities conducted in most schools, at least as perceived by faculty members.

4. The next logical step is to administer the 10-item aligned and balanced curriculum subscale in many districts that include different schools (e.g., elementary, middle, high schools) and schools whose status is known to be high-performing learning communities. A goal might be around 100 schools with a total faculty approaching 4,000, assuming high schools have somewhat larger faculties.

REFERENCES

Ceperley, P. E., & Squires, D. A. (2000). *Standard implementation indicators: Charting your course to high achievement*. Charleston, WV: AEL.

Cowley, K. S., & Meehan, M. L. (December 2002). *Student achievement gaps in high-performing schools: Differences as a function of the professional staff as a learning community.* Paper presented at the January 2003 Hawaii International Conference on Education, Waikiki, HI.

Elmore, R. (1995). Structural reform in educational practice. *Educational Researcher, 24*(9), 23-26.

Fullan, M. G. (1996). Turning systematic thinking on its head. *Phi Delta Kappan, 77*(6), 420-423.

Hargreaves, A. (1994). *Changing teachers, changing times.* London: Cassell.

Joyce, B. R. (1991). The doors to school improvement. *Educational Leadership, 14*(3), 10-12.

Meehan, M. L., Cowley, K. S., Craig, J. R., Balow, N. & Childers, R. D. (September 2002). *AEL continuous school improvement questionnaire user manual and technical report.* Charleston, WV: AEL.

Meehan, M. L., Cowley, K. S., Wiersma, W., Orletsky, S. R., Sattes, B. D., & Walsh, J. A. (July 2002). *Development and testing an instrument measuring a school faculty's commitment to continuous learning and improvement*. Paper presented at the 11th Annual CREATE National Evaluation Institute, Boise, ID.

Meehan, M. L., Wiersma, W., Cowley, K. S., Craig, J. R., Orletsky, S. R., & Childers, R. D. (July 2002). *School faculty as a high-performing learning community: Normative data from 132 schools*. Paper presented at the 11th annual CREATE national evaluation institute, Boise, ID.

Meehan, M. L., & Cowley, K. S. (January 2003). *A study of low-performing schools, high-performing schools, and high-performing learning communities*. Paper presented at the Hawaii international conference on education, Waikiki, HI.