Development of a Science Achievement Evaluation Instrument for a School Garden Program

C.D. Klemmer¹, T.M. Waliczek², and J.M. Zajicek³

ADDITIONAL INDEX WORDS. Bloom’s taxonomy, Texas Essential Knowledge and Skills (TEKS)

SUMMARY. School gardens show promise as a tool for developing science process skills through real-world investigations. However, little research data exist attesting to their actual effectiveness in enhancing students’ science achievement. The purpose of this study was to develop three cognitive test instruments for assessing science achievement gain of third, fourth, and fifth grade students using a garden curriculum. The development of the test instruments occurred in three phases: 1) an initial set of test instruments which served as a prototype for length, scope, and format; 2) an adapted set of test instruments which were piloted; and 3) a final set of test instruments which were used for the assessment of the school gardening curriculum. The final Cronbach’s alpha reliability for the final set of test questions was 0.82, indicating an acceptable level of internal consistency. Content validity of the test instruments developed for this study was established based on the science content standards specified in the Texas Essential Knowledge and Skills (TEKS) for each grade level along with the gardening curriculum, as well as the Science Scope and Sequence documents for Temple, Texas Independent School District (ISD). Construct validity was established for the testing instruments by soliciting help from various curriculum experts from the Temple ISD.

Acknowledgements: The authors express appreciation for the financial support of the Kellogg Foundation.

E valuation of educational programs is an important facet of the overall program because it allows the educator to determine if the program achieved the original objectives (Brown and Kiernan, 1998; Isaac and Michael, 1971). In addition, program evaluation helps provide evidence for educators of the overall value of the program in comparison to the investments. This can help to justify the delivery of the program over a period of time and during times of limited resources. It may also help to prioritize the programs that meet the needs of larger audiences (Voichick, 1991).

Gardening programs have been included in school curricula in various capacities for years (Bassett, 1979; Montessori, 1912). The popularity of using gardens to teach subjects such as science, math, nutrition and environmental education by incorporation of hands-on techniques has increased in the last several years (Eames-Sheavly, 1999; Hunken, 1994a, 1994b; Kiefer and Kemple, 1998; Lineberger and Zajicek, 2000; Moore and Wong, 1997; Shair, 1999; Skelly and Zajicek, 1998; Waliczek et al., 2003). State extension agencies and public gardens have developed educational guides to help teachers, administrators and parents learn how to incorporate school gardens into their own programs [Guy et al., 1996; Texas Agricultural Extension Service (TAEX), 1999a, 1999b].

The National Science Education Standards (NSES) establish common guidelines for science curriculum content, teaching, and assessment across the United States. Concern and debate over the need to improve science literacy led to the establishment of these standards (National Research Council, 2000). The curriculum standards included as part of the NSES determined content with which students should be familiar in order to function as scientifically literate adults. These standards were subsequently used to develop state curriculum guidelines. In Texas, these standards are the Texas Essential Knowledge and Skills, or TEKS, which were implemented in 1998. The TEKS specify content that
students should know at each grade level for each subject [Texas Education Agency (TEA), 2004].

School gardens use methods advocated by the National Science Education Standards and the TEKS for the teaching of science and reflect the constructivist philosophy of learning. Science is inherently active, and a sound science education program engages students as active participants involved in all the processes of science such as making predictions, planning and conducting experiments, observing phenomenon, collecting and assessing data, drawing conclusions, and developing new questions. Constructivist philosophies advocate student-centered learning through the use of a variety of learning styles, which allow students to build, or construct, their own knowledge for themselves based on their own personal context (Cole and Cole, 1996; Fogarty, 1999; Holloway, 1999; Howe and Jones, 1993; Mintzes et al., 1998; Perkins, 1999; Shapiro, 1994; Tobin, 1993).

Evaluation as part of garden program curriculum integration is a critical part of producing educational programs that demonstrate the impact of the program (Brown and Kiernan, 1998). At the time this study was conducted in 2000, there was no test instrument available to assess science achievement based on mastery of the curriculum standards set through the TEKS. Testing for achievement on the science TEKS was not implemented until Spring 2003, and test instruments were still in development at the time of this study (TEA, 2004). Therefore, cognitive test instruments measuring science achievement based on the TEKS science standards needed to be developed for third, fourth, and fifth grades for this study.

The purpose of this study was to develop three cognitive test instruments for assessing science achievement gain of third, fourth, and fifth grade students using a garden curriculum (Klemmer et al., 2005). Specific objectives included describing the constructs used to write the instrument, establishing face and content validity and testing the reliability for the instrument.

Materials and methods

Curriculum development. In order to evaluate the effect of a school garden program on students’ science achievement, the gardening program used needed to be similar in scope and content across all classrooms participating in the study. In order to maintain consistency between the schools and classrooms participating in the study, the same garden-based curriculum was implemented at all participating schools. The curriculum used for this study was a youth gardening curriculum developed by TAEX. The gardening curriculum is intended to educate youth about horticulture, health, nutrition, environmental science, and leadership (TAEX, 1999a, 1999b).

Although participants do not necessarily proceed through the program sequentially, and each chapter may be used independently, the chapters can be divided into three main groups with regard to their scope and content. The first four chapters pertain more directly to basic science concepts, and form a core foundation of horticulture-ecology-related learning for the program. The next three chapters reflect more applied science concepts, and consist of more subject-specific, thematic content. The last chapter is dedicated specifically to leadership and career development. The first four chapters—the core science content—were used for the purposes of this study (TAEX, 1999a, 1999b).

Development of the test instruments. The development of the test instruments occurred in three phases including the development of an initial set of test instruments which served as a prototype for length, scope, and format, an adapted set of test instruments which were piloted, and a final set of test instruments which were used for the assessment of the gardening curriculum. The steps of test development specified in Gall et al. (1996) were used as guidelines.

Establishment of content and construct validity. Content validity is the degree that the scores yielded by a test adequately represent the content that these scores were intended and designed to measure (Gall et al., 1996). The content validity of the test instruments developed for this study was established through two methods. First, the development of test items for each test instrument was directly based on the science content standards specified in the TEKS for each grade level along with the gardening curriculum, as well as the Science Scope and Sequence documents for Temple ISD, Temple, Texas (Table 1). The development of test items within this framework should, therefore, have strong content validity, since the items were directly based on content that the schools in the study were mandated to cover.

Secondly, the test items were developed with the input of the Science and Math Curriculum Coordinator for Temple ISD, a teacher and science specialist from one of the elementary schools involved in the study, and the Curriculum Coordinator for the gardening program, who was also formerly a third grade teacher. Each reviewed the instruments and made recommendations on minor changes that improved the clarity of several test items.

Construct validity is the extent to which a particular test can be shown to assess the construct that it purports to measure (Gall et al., 1996). The test instruments developed for this study were intended to measure science achievement, which is a cognitive trait and therefore difficult to measure. However, the TEKS on which the test instruments were based were themselves developed by the TEA over a period of several years, with the assistance of many science curriculum and evaluation specialists. They are comprised of content that the state education agency uses to define science achievement at each grade level. The TEKS can be determined, therefore, to have a fair level of construct validity for science achievement as defined by the state of Texas. This, then, established indirect construct validity for the instruments developed.

Initial test item development. With input from teachers and curriculum specialists, it was decided that each test instrument would be comprised of 40 items. This would allow an adequate number of items to ensure a strong measure of reliability, while still ensuring that the number of items would not be beyond the scope of elementary students’ ability. In addition, this length allowed for five items to be derived from each of the eight chapters in the gardening curriculum.

A draft outline of test items was developed by examining the science TEKS for each grade, along with the gardening curriculum concepts for each chapter, and the Science Scope and Sequence plan for Temple ISD.
Table 1. Multiple-choice statement examples and answers from the study of science achievement with school gardening (Klemmer et al., 2005) and the corresponding Texas Essential Knowledge and Skills (TEKS) objectives and school gardening curriculum content with which the statement and answer relate.

<table>
<thead>
<tr>
<th>Multiple-choice statement/question and corresponding answers</th>
<th>TEKS’ objectives</th>
<th>School gardening curriculum content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil is made up of rock particles, minerals, and decayed ________ and animal material</td>
<td>As students learn science skills, they identify components and processes of the natural world including properties of soil, effects of the oceans on land, and the role of the Sun as our major source of energy. In addition students identify the physical properties of matter and observe the addition or reduction of heat as an example of what can cause changes in states of matter.</td>
<td>Touchy-feely: To understand soil texture and the properties of different soil types and soil particles.</td>
</tr>
<tr>
<td>A. rock</td>
<td>Students learn the roles of living and nonliving components of simple systems and investigate differences between learned characteristics and inherited traits.</td>
<td>Mud Pies: To feel the difference in soil textures.</td>
</tr>
<tr>
<td>B. dirt</td>
<td>A system is a collection of cycles, structure and processes that interact. Students should understand a whole in terms of its components and how these components relate to each other and to the whole.</td>
<td>Shake, Rattle and Roll: To identify amounts of soil particles that make up a soil’s texture.</td>
</tr>
<tr>
<td>C. mud</td>
<td></td>
<td>Candy Aggregate: To create an edible model illustrating that soil is made up of many different components.</td>
</tr>
<tr>
<td>D. plant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A model representing all the parts of a plant would need to include ________.</td>
<td>The student knows that the natural world includes earth materials and objects in the sky.</td>
<td>Plant Parts Rap: To gain understanding of the main parts of a plant and role each performs.</td>
</tr>
<tr>
<td>A. roots, stems, leaves, flowers, fruit, and seeds.</td>
<td>A system is a collection of cycles, structure and processes that interact. Students should understand a whole in terms of its components and how these components relate to each other and to the whole.</td>
<td>Touch and Tell: To use information learned about plant parts to identify plant materials by touch.</td>
</tr>
<tr>
<td>B. roots, stems, leaves, tubers, fruit, and seeds.</td>
<td></td>
<td>Plant Parts We Eat: To identify the various plant parts that are used for food.</td>
</tr>
<tr>
<td>C. roots, petioles, leaves, flowers, fruit, and seeds.</td>
<td></td>
<td>Flower Dissection: To identify the different parts of a flower.</td>
</tr>
<tr>
<td>D. roots, stems, leaves, flowers, nodules, and seeds.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*TEKS specify content that students should know at each grade level for each subject.

Although test items were developed within the framework of subjects covered in the gardening curriculum, they were not labeled as such. The questions for each chapter did not relate to specific activities within the garden curriculum, but rather to the concepts taught through those activities. These concepts reflected the TEKS standards that all students in a particular grade were expected to know by the end of the school year (Table 1).

Developing test items that addressed 1) the subjects in the gardening curriculum chapters, 2) the science TEKS, and 3) the Temple Science Scope and Sequence plan proved exceedingly difficult. An inherent limitation of this initial plan was discovered in the lack of alignment between the science TEKS and the last four chapters of the gardening curriculum. These were correlated with the science TEKS, but were not as closely related to the basic science concepts as were those from the first four chapters. In addition, the Temple Science Scope and Sequence plan simply did not include many activities from these chapters. These last four chapters seemed somewhat outside the realm of the core curricular requirements as specified by the science TEKS. Because of the difficulties encountered, the test scope and outline was modified to include ten items from each of the first four chapters of the gardening curriculum.

**Correlation of test items to learning theory.** The rationale for test item development was based on Bloom’s taxonomy of knowledge (Bloom et al., 1956; Hopkins and Stanley, 1981). According to Bloom’s theory, mastery of one level of learning is necessary before advancing to the next level; therefore, learning is achieved through a hierarchical progression of levels of knowledge from simple to more complex (Fig. 1).

The 10 questions from each chapter were scaled in difficulty to assess a range of learning according to the first three levels of Bloom’s Taxonomy (Fig. 1).

Questions one through four for each chapter assessed students’ knowledge and comprehension of basic concepts, reflecting a competence in the lower two levels of Bloom’s Taxonomy. Questions five through eight for each chapter assessed students’ ability to comprehend concepts at a more advanced level, or their application of basic concepts, reflecting a somewhat more advanced level of understanding according to the com-
progression in cognitive development 

basic understanding as well as a natural 
els, showing an increase in the level of 

higher from lower to higher grade lev-

these questions would be progressively 

expected that student achievement on 

test instruments for grades three and 

of some additional questions on the 

chapter, which were basic knowledge 
or comprehension questions, were kept 

the same for all three grades in order to 

provide a measure of common compari-

son amongst the grades. Due to overlap 
of curriculum standards from one grade 

to the next, called vertical alignment, 

there was also overlap and repetition of 

some additional questions on the 

test instruments for grades three and 

and four, for grades four and five. It was 

expected that student achievement on 

these questions would be progressively 

higher from lower to higher grade lev-

els, showing an increase in the level of 

basic understanding as well as a natural 

progression in cognitive development and reading ability. 

The final two questions for each 

unit, which were higher level according 

to Bloom’s Taxonomy, assisted in dis-

criminating between lower and higher 

levels of learning. They also provided 

some measure as to whether students 

were able to transfer concepts learned 
to solve problems in new situations. 

If students were able to answer these 

more advanced questions correctly, a 

higher level of learning would be 

indicated. 

Following development of the 

40 test items, the instruments were 

reviewed by the Science and Math Cur-

riculum Coordinator for Temple ISD and by the Curriculum Coordinator 

for the gardening program. Finally, the 

reading level of each test instrument 

was checked using the Flesch-Kincaid 

grade level assessment as included in 

Microsoft Word 97 SR-1 (Microsoft, Redmond, Wash.). The reading 

level for the grade three instrument was 

rated a 3.5 grade level, while that of 

the grade four instrument was rated 

at a 4.5 grade level, and the grade 

five instrument was rated at a 4.7 

grade level. 

Format of the test instru-

ments. The tests developed were tra-

ditional paper and pencil tests, using a 
multiple-choice answer format (Table 

1). This format was chosen because it 

most closely conforms to current stan-

dardized tests used by most schools. 

Students selected the response that they 

considered correct, and then filled in 

the bubble beside that response. 

Pilot test. The pilot sample 

was drawn from a fourth grade class 

with 15 students. Piloting the fourth 

grade instrument allowed for the most 

overlap between the third and fifth 

grade tests since the first two questions 

from each chapter were the same on all 

three tests. Therefore, 20% of the test 

questions for all grades were piloted in 

one grade. In addition, there was some 

additional repetition of test questions 

between the third and fourth grade 

tests, and between the fourth and fifth 

grade tests. This was due to partial 

overlap of curriculum standards from 

one grade to the next to ensure vertical 

alignment between grade levels. 

Reliability of test instru-

ments. Reliability is the extent to 

which the same test scores would be 

obtained if the test were administered 

again (Gall et al., 1996), and ranges 

from –1.0 to +1.0. Test scores that 
yield a reliability of 0.80 or higher are 
gen erally considered reliable, although 

higher reliabilities of 0.90 or better are 

expected for scores from large-scale 

standardized cognitive tests (Gall et 

al., 1996). The most commonly used 

measure of reliability is Cronbach’s 

coefficient alpha. 

Pilot test reliability. Student 

responses on the pilot test were scored 

by hand and entered into an SPSS data 
sheet (version 10.0 for Windows95; 

SPSS, Chicago). The pilot data were 

then subjected to an inter-item reli-

ability analysis, yielding a Cronbach’s 

alpha coefficient of 0.40, and indicating a 

weakly reliable test instrument. 

Pilot test revisions for im-

proved reliability. Subsequent 

reliability analysis was conducted to 
determine which test items were weak-

est, and therefore requiring review and modification. An “alpha if item 
deleted” reliability analysis was 

carried out to obtain a series of Cronbach’s Alpha reliability coefficients for test 
scores when each test item was deleted. Items that increased the reliability 

coefficient above 0.44 when removed 

were indicated. These were deleted and the analysis rerun to obtain reli-

ability coefficients with the remaining 

items. Seven items were removed which 

resulted in an increased Cronbach’s Alpha reliability of 0.71. This process of 

removing the weakest contributing 

items was repeated through several 

iterations until the number of items 

remaining in the analysis was reduced 

to 23. This increased the Cronbach’s alpha reliability of the remaining items 

to 0.82. The 17 items removed to ob-

tain this reliability were thus deemed
Table 2. Cronbach’s alpha reliability factors* for pilot test scores when weak items are removed from the analysis in the study of science achievement with school gardening (Klemmer et al., 2005).

<table>
<thead>
<tr>
<th>Item numbers removed</th>
<th>Test questions (no.)</th>
<th>Cronbach’s alpha*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall reliability of pilot test scores</td>
<td>No items removed</td>
<td>40</td>
</tr>
<tr>
<td>Reliability of test scores with items removed</td>
<td>4, 6, 7, 19, 32, 34</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>4, 6, 7, 9, 11, 13, 14, 21, 30, 32, 34, 36</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>4, 5, 6, 7, 9, 11, 13, 15, 19, 21, 23, 30, 32, 34, 36</td>
<td>23</td>
</tr>
</tbody>
</table>

*Reliability or alpha scores range from –1.0 to +1.0, with higher scores indicating more reliability.

Table 3. Cronbach’s alpha reliability factor* for final test instruments in the study of science achievement with school gardening (Klemmer et al., 2005).

<table>
<thead>
<tr>
<th>Grade</th>
<th>Sample size (no.)</th>
<th>Test questions (no.)</th>
<th>Cronbach’s alpha*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall scores (grades 3, 4, and 5 combined)</td>
<td>647</td>
<td>40</td>
<td>0.86</td>
</tr>
<tr>
<td>Grade 3 scores</td>
<td>164</td>
<td>40</td>
<td>0.78</td>
</tr>
<tr>
<td>Grade 4 scores</td>
<td>253</td>
<td>40</td>
<td>0.78</td>
</tr>
<tr>
<td>Grade 5 scores</td>
<td>228</td>
<td>40</td>
<td>0.92</td>
</tr>
</tbody>
</table>

*Reliability or alpha scores range from –1.0 to +1.0, with higher scores indicating more reliability.

the weakest items (Table 2). These seventeen items were not permanently removed from the test; rather, they were reviewed for content and modified to improve misleading or confusing words or phrases. After reviewing the pilot test for item reliability analysis and item content, the third and fifth grade tests were adapted accordingly, since many of the questions were the same or similar.

Results and discussion

The final test instruments were administered to third, fourth, and fifth grade students in the Temple ISD in central Texas. The final sample consisted of a total of 647 students from seven different schools.

Reliability statistics for the final test instrument. A Cronbach’s alpha coefficient determined the reliability of the test scores for each grade level (Table 3). The overall reliability yielded a Cronbach’s alpha of 0.86. The reliability coefficient of 0.92 from the fifth grade test scores is highest, suggesting that the fifth grade test instrument is the strongest instrument. This may be due to an increased reading ability of older students, or this may reflect an increased aptitude for science in older students based on a higher level of cognitive development. The third and fourth grade test instrument reliability were 0.78 and 0.78, respectively (Table 2), still indicating an acceptable level of internal consistency (Gall et al., 1996; Sapp and Jensen, 1997).

Conclusions

Many curriculum programs and activity guides are available to enhance teachers’ ability to use the garden as a learning context for life sciences including: Botany for all Ages and Ecology for All Ages (Hunken, 1999a), Digging Deeper (Kiefer and Kemple, 1998), Lifelab (Center for Ecoliteracy, 1997), GrowLab (National Gardening Assn., 2003), Natural Learning (Moore and Wong, 1997), Project Learning Tree (American Forest Foundation, 2004), and numerous others. However, teachers today are being evaluated more than ever based on their students’ achievement. It is important for educators to not only have effective curriculum materials, but those that have been tested and shown to work, and correlated to mandated school curricular requirements.

In educational assessment, research and evaluation is necessary in order to make good choices between alternative programs or curricula. However, a reliable and valid tool to measure the variable of interest is not always available. The development of a new instrument is tedious, but worthwhile in order for good decision-making to occur. A good test instrument and research will provide objective evaluations and tangible scores. Collecting data on programs helps to ensure the use of effective practices, rather than faddish innovations. This type of program evaluation is especially important in this time of cost effectiveness in educational institutions in order to justify the incorporation of the programs with the most impact considering their use of resources.

Literature cited


