Classroom Observation Scale-Revised

User's Manual

Classroom Observation Scales Development Team:
Joyce VanTassel-Baska, Ed.D.
Linda Avery, Ph.D.
Jeanne Struck, Ph.D.
Annie Feng, Ed.D.
Bruce A. Bracken, Ph.D.
Diann Drummond, M.Ed.
Tamra Stambaugh, M.Ed.

Manual Development Team:
Joyce VanTassel-Baska
Chwee Quek
Annie Feng

The College of William and Mary
School of Education
Center for Gifted Education

2005

CHAPTER ONE

Introduction

Rationale for the COS-R

Over the last several years, there has been considerable evidence from different areas suggesting that how teachers behave in the classroom and the instructional approaches they employ significantly affect the degree that students learn. Sanders and Rivers (1996) have reported that the effects of ineffective teachers over three years as having a depressed effect on student achievement in math by as much as 54% regardless of the ability of the learner. Wenglinsky (2000) has found positive effects of using key practices such as critical thinking and metacognition on student learning in math and science across elementary and middle school levels. The literature in gifted education suggests that teacher behavior is the link to differentiated programs and services for this special population. Studies consistently suggest that regular classrooms have very limited differentiated activities (see Westberg, Archambault, Dobyns, & Salvin, 1993; Westberg & Daoust, 2003). Moreover gifted and talented teacher behaviors are not systematically monitored.

Classroom reform is highly dependent on positive teacher behavioral change in key areas. A study of math and science programs found that teachers will use strategies linked to content that show results with students (Kennedy, 1999). Collegiality and support have also been found to be necessary for change to occur (Garet, Porter, Desimone, Birman, & Yoon, 2001). Another study has shown that the use of higher-level reform behavior takes two or more years of intensive training to demonstrate results (Borko, Mayfield, Marion, Flexer & Cumbo, 1993). Based on these findings, it is clear that attention to classroom level instruction must be carefully monitored for the results employed to improve teaching.

Yet use of teacher evaluation alone or disconnected from on-going professional development appears to be unsuccessful as a catalyst for teacher change. A recent study of teacher evaluation practices in three districts in Wisconsin (Kimball, 2002) found that few teachers reported substantial changes in their instructional practice as a result of their evaluation experiences, and the large majority of teachers did not see the evaluation process as an incentive to seek out professional development opportunities.

Gifted education practice has had an uneasy alliance with key facets of educational reform (VanTassel-Baska, 1993), supporting the need for challenging standards but questioning the impact of inclusion on talent development. To its credit, the field is recognized for advancing the introduction of innovative instructional practices into the classroom, such as inquiry learning, critical and creative thinking skills, higher-order questioning strategies, metacognition, and the use of rich and varied curricular materials, rather than sole reliance on textbooks (Tomlinson & Callahan, 1992). Most recently, the introduction of content-based curriculum tied to state and national standards and evaluated on student learning gains (VanTassel-Baska, Bass, Ries, Poland, & Avery, 1998; VanTassel-Baska, Zuo, Avery, & Little, 2002) has again positioned the field in the forefront of curriculum reform, with the emphasis on “going through the standards and not around them” to achieve instructional impact.

In spite of such advances, there is little evidence that gifted education programs systematically assess student gains using appropriate learning measures (Avery & VanTassel-Baska, 2002). In the absence of student impact data, evaluators must often rely on the quality of the instructional experience as a proxy method for investigating program effectiveness. In fact, teacher effectiveness has been shown to be the main determinant of student progress (Sanders & Horn, 1998). Classroom observation provides a nexus between the input variables of the teacher and his/her students and the process of instruction itself, a process that combines instructional intent (goals and objectives), curriculum resources and materials, instructional and assessment strategies, and classroom management skills within a delimited period of time. In meeting the needs of the gifted learner, the observer must focus on three dimensions of practice: good teaching in general, key elements of educational reform, and differentiation for high ability learners.

There are certain assumptions made about the teaching act that underlay the use of any observation tool, especially one like the COS-R which is extensive and in-depth in respect to behaviors. Teaching is a complex social activity requiring the capacity to split attention by student, by area of the room, and by activity. It requires making multiple decisions during a teaching episode by instructional regulation, by strategies, use of time, and lesson emphasis. Moreover, teachers must think in complex ways in order to implement many behaviors simultaneously. For example, one type of thinking that teachers must employ relates to following the lesson plan, responding to pacing, modulating student-teacher interaction and ensuring a variety of stimuli. A second type of thinking required of teachers is awareness of teaching to multiple objectives, such as teaching a concept at the same time as they are teaching content and skills, and at the same time they are attending to the teaching of group dynamics. A third way of thinking involves orchestrating feedback strategies and responses, flexibility in grouping patterns, and questioning and activity choices. Planning, monitoring and assessing both group and individual learning is another level of thinking that teachers must employ in *medias res.* Finally, teachers need to think about sequencing their lessons. How does the current lesson relate to yesterday's and how will it link to tomorrow's?

Teaching has traditionally been a solitary activity; thus teachers are not socialized to having external observations of their work, and oftentimes are uncomfortable with the observation process. Yet, improvement in teaching clearly requires a change in teacher behaviors that promotes learning in students. Such improvement appears to imply the use of higher order thinking, problem solving and metacognitive approaches. In order to ensure that teachers are employing such strategies, some form of monitoring teacher behaviors must occur.

Goals for the COS-R

Classroom observation is a seminal part of education reform. It affords an opportunity to access the actual instructional experience that is at the heart of teaching and learning. It provides a nexus between the input variables of the teacher and his/her students and the process of instruction itself, a process that combines instructional intent (goals and objectives), curriculum resources and materials, instructional strategies, and classroom management skills within a delimited unit of time. It is the one part of professional development that allows the critical pieces of teacher knowledge and skills to come together in an authentic opportunity to gain insight about the quality of the learning experiences that are delivered.

One way of thinking about classroom observation component is to see it as a performance-based assessment of the teacher within the context of the learning environment. It affords many of the features of performance-based assessment with the teacher, rather than the student, as the unit of focus. For instance, it is a relatively open-ended experience, with teachers exercising much control over the selection of the lesson to be taught. It allows for the demonstration of complex and higher-order behaviors, recognizing that good teaching derives from a sophisticated set of skills that unfold in an integrated way. It also allows for self-assessment, providing a metacognitive dimension to the experience. Most importantly, by using a structured form, it provides a benchmark against which the teaching process can be assessed based on expectations derived from best practice in a given field.

However, it is important to distinguish between program and teacher evaluation in terms of the parameters used in the classroom observation process. The focus of program evaluation targets the collective whole. Teacher behavior is sampled through the classroom observation process and the intent is to allow teachers to prepare for the observation period in order to reduce the level of threat that permeates this current school climate of accountability. By aggregating data across classrooms, a snap-shot of instructional practice is created that helps inform our understanding of program quality.

Direct observation is also an important component of a teacher (or staff) evaluation process, but deeper sampling of the individual being assessed must occur when used for this purpose. Also, teacher evaluation should involve some opportunity for unplanned visits to the classroom to assess the consistency of instructional performance. While the two uses of classroom observation (program and teacher evaluation) can complement one another, it is important to acknowledge that the limited sampling template used for program evaluation is not sufficient to make inferences regarding individual staff performance tied to retention or promotion decisions.

Applications of the COS-R

If classroom practice for the high ability learner requires "differentiated" services, then how is this criterion translated into expectations for teacher performance and accountability? What constitutes “practice” in classrooms serving gifted learners? The COS-R provides a mechanism for using quantitative parameters to answer these questions. We cannot afford to ignore teacher behaviors that theory and research have found contribute to improved teaching effectiveness. The COS-R is one way to assess individual teacher performance in response to high ability learners.

Research suggests that while teachers of the gifted appeared strong in many categories of good teaching, they fall short in many others that examine differentiation practices (VanTassel-Baska, 2004). For instance, higher level thinking models were less in use than might be expected in classrooms when gifted learners form a critical mass of the student body. Very little emphasis was placed on accelerative strategies in gifted programs. Rate differences between gifted learners were not routinely attended to by classroom teachers, even those trained in gifted education. Use of models that explore different types of thinking and problem-solving behavior was not routinely employed in gifted program classrooms. The most prevalent problem-solving behavior observed was brainstorming, many times used in isolation of more fulsome models that examined a broader array of problem-solving behavior. Little evidence of the use of metacognition in gifted classrooms was observed. Moreover, little emphasis of out-of-class learning was found in observed teachers of the gifted classrooms as judged by their lack of provision for extended activities. Individualizing for differences within the gifted population, critical thinking, and problem solving were found to be underutilized in most observed classrooms. Core teaching behaviors like lesson planning and clarity in providing directions were more routinely observed. Curriculum reform elements were in evidence unevenly, with greater prevalence of concept teaching than other behaviors.

Thus the COS-R provides direct evidence of the need for specific emphases in development programs and may be effectively employed as a planning tool for that purpose. Professional development plans must take into account the evolving skills of teachers such that workshops and seminars are more tightly focused on needed skills as opposed to desired areas of interests. Gifted programs need to consider the sophistication of using higher level skills effectively and the need for teachers to develop these skills over time through appropriate methods.

The COS-R may also prove useful in charting overall growth in desirable teacher behaviors by school or district as has been found in other checklists with self-monitoring capabilities (Hague & Walker, 1996). Diagnoses of individual teacher skills needed to work effectively with gifted learners can also easily occur through the use of the COS-R. Videotape analysis of teaching can employ self-assessment using the instrument as well as using external observers.

Teaching is a performance art. Therefore, the test of success is demonstrating mastery of relevant techniques. Use of monitoring approaches including coaching, structured observation, and videotape analysis all increase the chance that teachers will improve in important ways in behaviors that optimize learning for our best students.

CHAPTER TWO

Scale Development

The process of COS-R development has involved several stages and forms occurring over more than a decade. At the first stage, the project team reviewed extant literature on effective teaching, differentiated instruction for gifted learners, educational reform and change, and professional development research. Based on expectations derived from best practices in mainstream and gifted education classrooms, the form was developed to be utilized in all classrooms and in all subject areas. Only the most critical behaviors for general teaching and differentiation features culled from research-based evidence of effective classroom-based instructional behaviors were included. These focus on the use of strategies that promote student learning and growth especially in the area of higher-order thinking, problem-solving and metacognition. Clusters were created based on predominant teaching behaviors within each category. In its final iteration, the COS-R comprised a total of 25 items in six clusters. Table 1 shows the six clusters and the number of items in each.

Table 1 Number *of items per cluster*

	No. of items
General Teaching Behaviors	
Curriculum Planning and Delivery	5
Differentiated Teaching Behaviors	
Accommodations for Individual Differences	4
Problem Solving	3
Critical Thinking Strategies	4
Creative Thinking Strategies	4
Research Strategies	5

A description of the distinctive behaviors that characterize effective teachers' classes in each of these clusters follows.

General Teaching Behaviors

Curriculum Planning and Delivery (CPD)

Effective teachers thoroughly plan and organize for instruction (Shellard & Protheroe, 2000). In planning and structuring instruction and activities, teachers have an explicit set of high expectations for student performance, and communicate the importance of learning (Cawelti, 1999). They incorporate activities for students to apply new knowledge (Marzano, Pickering & McTighe, 1993), engage students in metacognitive processes (Costa, 2001; Wang, Haertel & Walberg, 1993), and encourage student reflection and expression of thought (Good & Brophy, 1997).

Differentiated Teaching Behaviors

Instructional practices in classrooms for the gifted emphasize the importance of concept development, thinking and reasoning, problem solving, and flexible accommodations for working with highly able learners (VanTassel-Baska, 2003). Gifted learners are motivated when engaged in learning basic skills in context rather than in isolation, functioning consistently at high levels of thinking, making connections among disciplines, solving real problems, presenting products to real audiences and dealing with ambiguities and behaving like professionals in the field (Tomlinson,1996).

Accommodations for Individual Differences (AID)

Gifted students are not homogeneous. They differ in their abilities, their readiness to learn, their interests, and their styles of learning. Effective teachers recognize student variance and address it appropriately. (Tomlinson,1999). Besides whole class instruction and discussion, teachers design structured activities and questions to allow students to discover ideas individually. They provide opportunities for individual or group learning and accommodate individual or subgroup differences by allowing choice in material and task selection (VanTassel-Baska, 2003).

Problem solving (PS)

Gifted students have early capacity for problem solving; and they possess unique characteristics that make them adept at problem finding (Gallagher, 2001). The items in this cluster are derived from the stages of well-established research-based models of problem solving like that of Parnes' and Osborne's model (Isakson, Puccio, & Treffinger, 1993).

Critical Thinking Strategies (CRI)

Effective teachers of the gifted emphasize higher order thinking skills that are integral to all content areas and in everyday life experiences (Paul, 1992). Students demonstrate their understanding of advanced content by making generalizations from the concrete to the abstract and vice-versa, and synthesizing information within and across disciplines. (VanTassel-Baska, 2003; Wenglinsky, 2000).

Creative Thinking Strategies (CRE)

Torrance (1981, cited in Cramond, 2001), found in a study that "teachers who made a difference were those who enabled their students to hold on to their creativity" (p. 405). Effective teachers use an array of strategies to foster creativity. They have students explore diverse view points to reframe ideas. They solicit diverse thoughts about issues from students and encourage them to elaborate on their ideas. Teachers demonstrate open-mindedness and tolerance of imaginative and unusual solutions to problems, and encourage students to do the same. (Davis, 2003; Treffinger,1995).

Research Strategies (RS)

Gifted students should have the opportunity for independent study and be equipped with the prerequisite skills for effective research, and develop these skills to a sophisticated level (Reis & Renzulli, 1992). Students should be required to use appropriate and varied research techniques to gather evidence from multiple sources, interpret, draw inferences and make conclusions from them. They also should be given the opportunity to communicate their research findings to relevant audiences (VanTassel-Baska & Little, 2003).

Examples of items in the clusters

As mentioned earlier, the COS-R is divided into six subscales. The first subscale focuses on *Curriculum Planning and Delivery*. Examples of observable teacher behaviors for this subscale include setting high expectations for student performance and asking students to reflect on what they have learned. The second subscale focuses on *Accommodations for Individual Differences*; examples include accommodating individual differences through materials, conferencing, and/or task assignments and encouraging multiple interpretations. The third subscale focuses on *Problem Solving*, specifically the heuristic of brainstorming, problem identification and definition, and developing problem-solutions based on generalizations. The fourth subscale targets *Critical Thinking Strategies* and includes skills such as engaging students in comparing and contrasting ideas as well as encouraging students to synthesize information within and across disciplines. The fifth subscale focuses on *Creative Thinking Strategies*; examples include soliciting diverse thoughts about issues or ideas from students and encouraging students to demonstrate open-mindedness and tolerance of imaginative solutions to problems. Finally, the sixth subscale focuses on *Research Strategies* such as gathering evidence from multiple sources, analyzing data, and encouraging students to identify consequences and implications of their findings.

Demographic information about the class observed is captured in the first section of the COS-R. Information about the teacher, as well as the class (grade level, number of boys and girls, ethnicity, classroom desk arrangement, etc.) provides the context for the lesson observation.

Subject-based Indicators

To complement the six clusters of teaching behaviors, sets of domain-specific indicators were developed. The purpose was to provide observers with illustrative examples of observable classroom behaviors as they pertain to different subject areas. This was to reduce variability in the interpretation of individual items within each cluster in any content area, and to enhance inter-rater reliability. For example, for the item “Encouraged students to judge or evaluate situations, problems, or issues” in the cluster of “Critical Thinking Strategies”, the observable evidence for each of the subject areas might be as follows:

Math	Asked boundary/condition questions about proof/theorem such as “Under what conditions will this proof hold up and under what conditions it will not?
Science	Asked questions such as: “Were the results replicable?” “Were the data reliable?” “Was the experiment well-designed?”
Literature	Encouraged students to form interpretative hypotheses and test them on further reading or subsequent readings by applying criteria of plausibility and consistency
Social Studies	Asked questions about the implications of context for understanding a primary source document
Foreign/Second Language	Asked questions about an author’s purpose and assumptions

The indicators for each subject area were developed for each behavioral item and reviewed by content specialists using content-based curricula tied to state and national standards, and feedback from content and gifted education specialists was sought to improve the clarity of indicators’ description and ensure that prototypical examples were included.   The list of behavioral indicators, which usually numbers 3-5, is not meant to be exhaustive, merely illustrative. Thus appropriate behaviors in each category that are not on the list should also be rated.

Rating Scale

Each item on the scale is rated for its level of effectiveness. There are three levels on this rating scale with a rubric description of each level:

3 Effective	The teacher evidenced careful planning and classroom flexibility in implementation of the behavior, eliciting many appropriate student responses. The teacher was clear, and sustained focus on the purposes of learning
2 Somewhat effective	The teacher evidenced some planning and/or classroom flexibility in implementation of the behavior, eliciting some appropriate student responses. The teacher was sometimes clear and focused on the purposes of learning
1 Ineffective	The teacher evidenced little or no planning and/or classroom flexibility in implementation of the behavior, eliciting minimal appropriate student responses. The teacher was unclear and unfocused regarding the purpose of learning.
N/O Not observed	The listed behavior was not demonstrated during the time of the observation.

Effectiveness issues

Observers consider the following guidelines when rating the effectiveness of a teaching behavior:

- An item checked "N/O" is not a negative rating and does not address teacher effectiveness in any way.
- An item checked "1" is more negative than an item that is checked "N/O".
- In order for a behavior to be rated, it has to be deliberate and sustained.
- Not all of the indicators for each item in a subscale need to be present to warrant a rating of 2 or even 3.
- Appropriate behaviors that are not listed in the descriptors should be acknowledged through the "comments" section.
- A rating of "3" (effective) is not equivalent to exemplary and should not be withheld on grounds that the observed behavior is not outstanding.

Scale Blueprint

Item development process

Items were developed to reflect the key behaviors in each cluster. The minimum number of behaviors per cluster was three, and the maximum was five. Items were written with the following considerations**:**

- All Items should be succinct.
- Use of jargon was avoided.
- All items were written in the active voice to clearly describe the teaching behavior.
- Only instructional behaviors that can be observed in classroom settings were included.
- Items were worded positively.

To emphasize the teacher as the unit of focus, all items begin with the stem "The teacher…" followed by the description of the key instructional behavior. Similarly, the subject-based descriptions of teaching behaviors are focused on the actions of the teacher. The items in each of the clusters are given in Table 2.

Table 2 Categories and items within each category

Categories	**Items**
Curriculum Planning and Delivery (CPD)	• set high expectations for student performance • incorporated activities for students to apply new knowledge • engaged students in planning, monitoring or assessing their learning • encouraged students to express their thoughts • had students reflect on what they had learned
Accommodations for Individual Differences (AID)	• provided opportunities for independent or group learning • accommodated individual or subgroup differences • encouraged multiple interpretations of events and situations • allowed students to discover key ideas individually
Problem solving (PS)	• employed brainstorming techniques • engaged students in problem identification and definition • engaged students in solution-finding activities and comprehensive solution articulation
Critical Thinking Strategies (CRI)	• encouraged students to judge or evaluate situations, problems, or issue • engaged students in comparing and contrasting ideas • provided opportunities for students to generalize from concrete information to the abstract • encouraged student synthesis or summary of information within or across disciplines.
Creative Thinking Strategies (CRE)	• solicited many diverse thoughts about issues or ideas • engaged students in the exploration of diverse points of view to reframe ideas • encouraged students to demonstrate open-mindedness and tolerance of imaginative, sometimes playful solutions to problems • provided opportunities for students to develop and elaborate on their ideas
Research Strategies (RS)	• required students to gather evidence from multiple sources through research-based techniques • provided opportunities for students to analyze data and represent it in appropriate charts, graphs, or tables • asked questions to assist students in making inferences from data and drawing conclusions • encouraged students to determine implications and consequences of findings • provided time for students to communicate research study findings to relevant audiences in a formal report/presentation.

Piloting the form

The original version of the COS-R was the Classroom Observation Form (COF) which had 40 items in nine sub-categories. No sub-category had fewer than three items, while the number of items in the largest sub-category was nine. The COF was piloted with 50 teachers teaching in the Saturday Enrichment Program at the College of William and Mary. Based on the data gathered from the pilot, items in each of the sub-categories were reviewed for their contributions to the overall reliability (i.e., internal consistency) of the instrument, while ensuring that they were representative of key teaching behaviors. To reduce redundancy according to the rule of parsimony, and to increase the rigor of the scale, the number of sub-categories and items were reduced. The final version of the COF, re-named the COS-R has 25 items in six sub-categories.

Data using the COS-R were collected in two waves of observation in the first two years of implementation of Project Athena, a research study funded by a United States Department of Education Javits grant to explore the effects of a language arts curriculum treatment on students' reading achievement and critical thinking and on teaching practices in Grades 3, 4, and 5. The COS-R was also used in a replication study with a group of teachers in an enrichment program at the College of William & Mary. Details of the pilot and replication study are discussed in Chapter Four.

CHAPTER THREE

Implementation and scoring

Observers are encouraged to use a script sheet to do a narrative recording of instructional behaviors observed during the lesson. To be included in the script are a description of teacher's actions, time spent on activities and transitions between activities, and a record of teachers' questions and student responses as well as the pattern of interaction. Observers are also encouraged to note questions they might have about the lesson.

Observers are to script their observations independently, and then code them according to the 25 items in the six clusters. Observations are typically based on a 30 to 50-minute lesson. Each observer is to determine first if a behavior is observed or not observed. If it not observed, the "N/O" column should be checked. If a behavior is observed, the observer is to rate its effectiveness, using the 3-point scale provided.

Scoring

Observers visit classrooms in pairs. Each observer uses the lesson script to complete the COS-R checklist individually. No items should be left blank. After each observer has completed the form, the team should confer to complete a consensus form. Observers are strongly encouraged to cite evidence in the lesson script to support their ratings, and to achieve consensus, using script-based evidence. Scores for each individual item as well as each subscale can be computed to provide subscale means and item means to provide a snap-shot of the nature of instructional practice across classrooms.

Recommendations for teaming

For any observation involving people, there is inevitably an element of subjectivity, even in the use of an 'objective' instrument. Observers come from different backgrounds which may color their perspective. There is subjectivity in interpreting the occurrence of a behavior in the classroom, and/or subjectivity in the interpretation of the item on the scale. One way to reduce the subjectivity is to have a pair of observers in the same classroom, and to have the pair reach consensus.

It is recommended that teaming of observers should take into consideration the following:

- There should be at least one expert in the team who has the experience and expertise to make valid judgments. Such experts may be content supervisors, principals, or coordinators of gifted programs. An expert/novice team is advantageous in that the novice could bring fresh insights to the observation and learn from the expert more about the application of teaching behaviors to classroom practice.
- Pair an internal with an external observer. The internal observer would be aware of factors in the school that might have impacted the lesson that the external

observer would not know of, while the external observer, unfamiliar with the school, might see things in a different light.

- It is useful to have a school administrator on the team in order to ensure the ongoing use of the form.

Whatever the basis for the team formation, the primary concern should be to ensure a holistic, balanced, and fair assessment of the lesson observed.

CHAPTER FOUR

Technical Adequacy

The technical characteristics of the COS-R were investigated during two waves of data collection during the first two years of the implementation of Project Athena.

Reliability

Reliability is the degree to which measures are free from error and therefore yield consistent results. There are three major sources of error: factors in the measure itself (poorly written items), factors in the people answering the items on the measure, and scoring factors (clarity of scoring rubrics) (Rudner & Schafer, 2001). Internal consistency refers to the extent to which all questions or items assess the same characteristic, skill, or quality. Wasserman & Bracken (2003) suggest that scales intended for research applications should minimally be reliable at a level of .70, and preferably .80. This recommended level guided the selection of items for the final version of the COS-R.

The COF was piloted on a group of 50 teachers teaching in the Saturday Enrichment Program for gifted children at the College of William and Mary. Based on the pilot data, the number of categories was reduced to six, and the number of items was reduced to 25. Table 3 presents these pilot data.

Table 3: *COS-R Reliability Table*

COS-R	Alpha
Curriculum Planning & Delivery	.54
Accommodations for Individual Differences	.69
Problem Solving	.73
Critical Thinking Strategies	.66
Creative Thinking Strategies	.63
Research Strategies	.89
Overall	**.92**

The COS-R was used twice in the Athena Classroom Observation during Fall 2003 and Spring 2004. Twenty three teams of observers visited 73 classrooms in two rounds of observations. Two waves of data were collected from these observations. Item reliability analyses were conducted for the overall scale, as well as each of the six subscales. Table 4 shows the results of these analyses.

Table 4 *COS-R Reliability from Two Observation Periods*

Scale	1st Observation (N=72)	2nd Observation (N=58-62)
Curriculum Planning & Delivery	.79	.67
Accommodations for Individual Differences	.68	.73
Problem Solving	.82	.94
Critical Thinking Strategies	.65	.78
Creative Thinking Strategies	.86	.77
Research Strategies	.83	.83
Overall COS-R Scale	**.91**	**.93**

The analyses of the two observation periods showed that overall, the scale was highly reliable (Alpha .91 to .93). For both observations, the subscale reliability for all the clusters averaged above .70. These high reliability coefficients across both observations attest to the reliability of the items on the instrument. The inter-rater reliability of the COS-R also reached .87 and .89, respectively across each observation period.

Inter-rater reliability

Inter-rater reliability ensures that there is consistency in scoring the COS-R, regardless of who the trained observer is in the classroom. In order to assess inter-rater reliability, the research team had each observer complete an individual rating for each observation in the pilot and then analyzed the correlation between each pair of ratings.

Training on the COS-R

In order to enhance consistency of interpretation, each potential user of the form receives a half day training on the form. This training consists of:

- Introduction to the COS-R
- Observation of video-taped lessons which are scripted and rated
- Discussion of ratings
- 2nd video tape practice session followed by discussion.

Content validity

The COS-R was sent to six experts in gifted education to review its content validity. On a prescribed form developed for this purpose, reviewers were asked to rate the COS-R on two dimensions:

- The importance of each behavioral item on the scale
  Raters were to use a 3-point scale to judge the level of importance, with 3 being 'very important' and 1 being 'not important'.
- The accuracy of the language used to describe the behavior

The clarity of the language was judged on a 3-point scale, with 3 being "very clear", and 1 designating "a lack of clarity" in the language used to describe the behavior.

Three professors and scholars in gifted education and three practitioners and administrators in local school districts were identified for the content validity exercise. Four of them returned the prescribed forms with their ratings and comments. Based on the agreement of these four experts, the intra-class coefficient Alpha was used to assess rater agreement on the scale. Their agreement on the two dimensions of the validity was .86 for the importance of item, and .99 for clarity of language. These results are presented in Table 5.

Table 5 *Content Validity Table*

Dimension	Intra-Class Coefficient
Importance of Instructional Behavioral Items in the Teacher Observation Scale	.86
Clarity of Language in Describing Items in the Teacher Observation Scale	.99
COS-R Validity Content	**.98**

The reliability study of the COS-R was replicated in Spring, 2004 with seventeen teachers in a Saturday Enrichment Program for gifted students at the College of William and Mary. Of the 17 classes taught by the participating teachers, five were science-related classes, five were humanities-related, three were math-related, and two focused on developing problem-solving skills. Two observers who had been trained and had experience using the COS-R observed the teachers in classes spanning grades pre-K to Grade 9. The inter-rater reliability for this replication study was .92. The sub-scale reliability for two of the clusters Curriculum Planning and Delivery (Alpha: .91) and Creative Thinking Skills (Alpha: .89) were strong.

Future validation exercises

The COS-R is being used in secondary classrooms in five different subject areas in another cultural setting. The data gathered will be analyzed and compared to existing data and reported in later drafts of this user's manual.

References

Avery, L. D., & VanTassel-Baska, J. (2002). The impact of gifted education evaluation at state and local levels: Translating results into action. *Journal for the Education of the Gifted*.

Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1993). Teachers developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education, 13*, 259-278Borko, Mayfield, Marion, Flexer & Cumbo (1993).

Cawelti, G. (Ed.). (1999). Handbook *of research on improving student achievement: The schooling practices that matter most. 2nd edition.* Arlington, VA: Educational Research Service.

Costa, A.L.(Ed.). (2001). *Developing minds: A resource book for teaching thinking 3rd edition.* Alexandria, VA: Association for Supervision and Curriculum Development

Cramond, B. (2001). Fostering Creative thinking. In F. Karnes, & S. Bean, Suzanne (Eds.) *Methods and Materials for Teaching the Gifted*. (pp.399-444). Tx: Prufrock Press

Garet, M.S., Porter, A.C., Desimone, L. Birman, B.F. & Yoon, K.S. (2001). What makes professional development effective: Results from a national sample of teachers. *American Research Journal, 38,* 915-945.

Good, T.L. & Brophy, J.E. (1997). *Looking in classrooms, 7th edition..Chapter 2*: *Increasing teacher awareness through observation.* New York: Longman

Gusky, T.R. (2003). *Evaluating professional development*. Thousand Oaks, CA: Corwin Press.

Hague, S. A., & Walker, C. (1996). *Creating powerful learning opportunities for all children: The development and use of a self-monitoring checklist for teachers.* ERIC document ED396844

Isakson, S. G., Puccio, G.J., & Treffinger, D.J. (1993). An ecological approach top creativity research: Profiling for creative problem solving. *Journal of Creative Behavior, 27,* 149-170.

Kennedy, M. (1999). Form and substance in mathematics and science professional development. *NISE Brief, 3 (2),* 1-7.

Kimball, S.M. (2002). *Analysis of feedback, enabling conditions, and fairness perceptions.* University of Wisconsin, Wisconsin Center for Education Research, Consortium for Policy Research in Education.

Little, Feng, VanTassel-Baska, Rogers & Avery (2002). *Project Phoenix: Final report on social studies curriculum effectiveness.* Williamsburg, VA: Center for Gifted Education. The College of William and Mary.

Marzano, R. J., Gaddy, B. B., & Dean, C. (2000). *What works in classroom instruction.* Aurora, CO: Mid-continent Research for Education and Learning.

Marzano, R. J. (1998). *A theory-based meta-analysis of research on instruction.* Aurora, CO: Mid-continent Research for Education and Learning

Marzano, R. J.,Pickering, D. & McTighe, J. (1993). *Assessing Student Outcomes*. Alexandria: Association for Supervision and Curriculum Development, v, 1-5, 37-43.

Paul, R. (1992). *Critical thinking: What every person needs to survive in a rapidly changing world*. Rohnert Park, CA: Center for Critical Thinking and Moral Critique, Sonoma State University.

Reis, S. M., & Renzulli, J. S. (1992). The library media specialist's role in teaching independent study skills to high ability students. Library Media Quarterly, 21, 27-35.

Rudner & Schafer, (2001). *Reliability.* ERIC Clearinghouse on Assessment and Evaluation College Park MD, ED 458213. http://www.ericdigests.org/2002-2/reliability.htm

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessment system (TVASS) database: Implications for educational research and evaluation. *Journal of Personnel Evaluation in Education, 12 (3),* 247-189.

Sanders, W. I. & Rivers, J.C. (1996). *Cumulative and residual effects of teachers on future students' academic achievement.* Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Shellard, E. & Protheroe, N. (2000). Effective teaching: How do we know it when we see it? *The Informed Educator Series*. Arlington, BA: Education Research Service

Tomlinson, C. A. (1999). *The differentiated classroom: Responding to the needs of all learners.* Alexandria, VA: Association for Supervision and Curriculum Development.

Tomlinson, C. A., & Callahan, C. M. (1992). Contributions of gifted education to general education in a time of change. *Gifted Child Quarterly, 36,* 183-189.

Treffinger, D. (1995). *Creativity, creative thinking, and critical thinking. In search of definitions.* Sarasota: Center for Creative Thinking.

VanTassel-Baska, J. (1993). Linking curriculum development for the gifted to school reform and restructuring. *Gifted Child Quarterly, 1,* 34-37.

VanTassel-Baska, J. (2003). *Curriculum Planning and Instructional Design for Gifted Learners.* Denver, CO: Love Publishing.

VanTassel_Baska, J. (2004). Assessing classroom practice: The use of a structured observation form. In J VanTassel-Baska & A. Feng (Eds.) *Designing and utilizing evaluation for gifted program improvement.* (pp. 87-108). Tx: Prufrock Press

Van Tassel-Baska, J. & Little, C. (Eds.). (2003). *Content based curriculum for high ability learners.* Waco: Tx: Prufrock Press

VanTassel-Baska, J., Bass, G., Ries, R., Poland, D. & Avery, L.D. (1998). A national pilot study of science curriculum effectiveness for high ability learners. *Gifted Child Quarterly, 42,* 200-211.

VanTassel-Baska,J., Zuo, L., Avery, L. D, & Little, C.A. (2002). A curriculum study of gifted student learning in the language arts. *Gifted Child Quarterly, 46,* 30-44.

Wang, Haertel & Walberg, (1993). What helps students learn? *Educational Leadership,* 74-79.

Wasserman, J. D., & Bracken, B. A. (2003). Psychometric considerations of assessment procedures. In J. Graham and J. Naglieri (Eds*). Handbook of assessment psychology (pp. 43 – 66).* New York: Wiley.

Wenglinsky, H. (2000). *How teaching matters.* Princeton, NY: Educational Testing Service.

Westberg, K, Archambault, F., Dobyns & Salvin, T. (1993). The classroom practices observation study. *Journal for the Education of the Gifted, 16,* 120-146.

Westberg, K. & Daoust, (2003, Fall). The Results of the Replication of the Classroom Practices Survey Replication in Two States. *The National Research Center on Gifted and Talented Newsletter,* 3-8.