Journal of Classroom Interaction, ISSN 0749-4025. Copyright © 2005, Vol. 40.2, pages 44–66.

Drawing from Teacher Effectivess Research and Research into Teacher Interpersonal Behaviour to Establish a Teacher Evaluation System: A Study on the Use of Student Ratings to Evaluate Teacher Behaviour

Leonidas Kyriakides
DEPARTMENT OF EDUCATION, UNIVERSITY OF CYPRUS, USA

ABSTRACT

This paper presents results of a study attempting to identify the extent to which teacher effectiveness research and research into teacher interpersonal behavior can help us collect valid and reliable evaluative data from students about their teacher behavior. The major findings of teacher effectiveness research are outlined and the process that was followed in order to design questionnaire measuring student views of their teacher behavior in the classroom is presented. The main findings of research into teacher interpersonal behavior are also presented, and the process that was followed in order to translate, to Greek, the Questionnaire on Teacher Interaction (QTI) and examine its content validity is described. A stratified sample of 38 primary schools in Cyprus was selected and the two questionnaires were administered to all year-6 students (N=1973) from each class (N=92) of the school sample. Evidence supporting the reliability, discriminate validity, and construct validity of each questionnaire is provided. Data collected from most of the scales of each questionnaire were associated with student achievement gains in both cognitive (Mathematics and Greek Language) and affective outcomes of schooling. Implications of findings for the development of a teacher evaluation system based on the main findings of teacher effectiveness research and research into teacher interpersonal behavior are drawn. Suggestions for further research are also provided.

INTRODUCTION

One of the basic problems that deplore most educational systems concerns the need of developing a valid personnel evaluation system in order to measure teacher performance and to contribute in their professional development. Ellett and Garland (1987) studied teacher evaluation practices in the 100 largest school districts in the USA and found some key problems. These appear to be longstanding because a follow-up study 10 years later found little change (Loup, Garland, Ellett, & Rugutt, 1997). A similar situation to the one described in the USA can be identified in many other countries, and especially in countries where teacher evaluation systems are expected to achieve summative purposes (Kyriakides & Campbell, 2003). The difficulties that policy makers face in their attempt to develop valid evaluation systems partly arise from the fact that teacher evaluators are expected to justify the quality of teaching and the effectiveness of teachers and schools (Stronge & Ostrander, 1997) while it is widely accepted that teaching is a complex act that occurs in many forms and contexts thereby its quality should be looked in a variety of ways (McGreal, 1983). The previous argument leads evaluators to the need of using multiple sources of information about teachers' role to make the best personnel decisions; such as the outcomes of their students, external observations, views of their colleagues, and client surveys.

However, the teacher evaluation methods used in most countries are based on a model that requires administrators/ external evaluators to diagnose weaknesses and subsequently to prescribe solutions. Although classroom observation can be a meaningful and vital aspect of a comprehensive teacher evaluation system, it has major drawbacks as a single-source methodology (Stronge & Ostrander, 1997). Well-designed empirical studies depict administrators as inaccurate raters of teacher performance, because of the artificial nature of scheduled observations, the failure to reflect teacher responsibilities outside the classroom, the infrequency of observations, the fact that only a portion of the full repertoire of teacher duties and responsibilities can be observed in any one observation (Stronge, Helm, & Tucker, 1995), and the low correlation of administrator ratings with data gained from other sources

(Peterson, 1987). In addition to empirical studies that show the statistical inaccuracy of administrator ratings, survey studies of teachers and administrators indicate extremely low levels of respect for the procedures within the profession (Peterson 1995). Therefore, additional empirical evidence is needed to identify how different sources of data can be used in order to collect valid data on the quality of teaching. In this context, this paper is an attempt to provide suggestions on how the quality of data that can be collected from student ratings of teacher behavior in the classroom can be improved. Thus, the following section presents a brief review of literature concerning student ratings of teacher performance and argues that student reports should be considered as an important part of measuring teacher behavior.

STUDENT RATINGS OF TEACHER PERFORMANCE

Student ratings of instruction are probably the most common method used in higher education to assess instruction (Fresko & Nasser, 2001); a trend that is likely to continue with the increased emphasis on accountability. Student ratings of teacher performance are used consistently in higher education but not without criticism. They constitute the most controversial type of client feedback (CEPI, 2000) because there are many concerns regarding their reliability and validity. Despite these concerns, the extensive research covered in numerous works converge that there are several strong arguments for using student ratings as a source for evaluating teachers. As direct recipients of the teaching-learning process, students are in a key position to provide information about teacher effectiveness. Specifically, Stronge and Ostrander (1997) argue that "students are the only ones among all the teacher's clients, who have direct knowledge about classroom practices on regular basis" (p. 145). Moreover, student ratings constitute a main source of information regarding the development of motivation in the classroom, opportunity for learning, degree of rapport and communication developed between teacher and student, and classroom equity (Aleamoni, 1981). In addition, students are good sources of information about their instructors because they know their own situation well, have closely and recently observed a number of teachers, uniquely know how students think and feel, and directly benefit from good teaching (Peterson, 1987). Along with the previously mentioned reasons, student reports are defensible sources of information about teacher performance because of the fact that they represent participation in a process often characterized as "democratic decision making" (Scriven, 1995).

In spite of the widespread use and reliance on student ratings in higher education, Aleamoni (1999) points out that "they remain suspect as means of evaluating instructional effectiveness" (p. 153). The main objections to the use of student ratings are related to their reliability and validity, especially in the case of using student ratings for summative purposes. A meta-analysis of 38 studies on the quality of student ratings conducted recently (Kyriakides, 2001) revealed that the great majority of the studies were concerned with the use of student ratings in higher education. This seems to provide further support to the argument that although it is accepted that there are several strong reasons for using student ratings to evaluate teachers, still not much effort has gone into the development of principles and practices of this source of data at the K-12 level (Peterson, Wahlquist, & Bone, 2000). Moreover, the great majority of these 38 studies was focused on investigating factors affecting students' ratings of the effectiveness of their instructors of higher education. Although such research is useful for investigating the validity of student ratings, almost no emphasis on the content of the student questionnaires used to measure teacher effectiveness is given. As a consequence, only few studies deal with the theoretical background upon which student questionnaires should be based and examine the construct validity of the questionnaires (e.g., de Jong & Westerhof, 2001; Marsh & Roche, 1997). Furthermore, it was not possible to identify studies that attempted to use different methodological approaches to evaluate the various forms of reliability and validity of student ratings. It should, however, be acknowledged that researchers investigating the reliability of student ratings attempted to investigate the stability of student ratings across time, across courses, and across instructors (Young, Delli, & Johnson, 1999). Although considerable literature questions the reliability of student ratings, recent research indicates just the opposite. The stability of student ratings from one year to the next resulted in substantial correlations in the range of 0.87 to 0.89. Moreover, the correlation between student ratings of the same instructors and courses ranged from 0.70 to 0.87 (Kyriakides, 2001).

Validity of student ratings is a critically important issue with far-reaching implications for using student ratings to measure teacher behavior. The term "validity" denotes the scientific utility of a measuring instrument, broadly statable in terms of how well it measures what it purports to measure (Nunnally & Bernstein, 1994). It is therefore important to specify how well the instrument has met the standards by which it is judged. Moreover, Sax (1997) claims that validity is defined as the extent to which measurements are useful in

making decisions and providing explanations relevant to a given purpose. To the extent that measurements fail to improve effective decision-making by providing misleading or irrelevant information, they are invalid. No matter how reliable they are, measurements lack utility if they are not valid for some desired purpose. Therefore, researchers should not only attempt to investigate the reliability of instruments used to measure teacher performance through student views. It is also important to deal with the process of designing such instruments and investigate the construct validity of the proposed instruments.

A test's construct validity is defined by the extent to which a set of items measures the theoretical construct it was designed to measure (Allen & Yen, 1979). Construct validity is an ongoing process whereby a test is evaluated in the light of a specific construct. Thus, construct validity is evaluated in the context of a set of hypotheses and the assessed validity of a test rests within the domain of these hypotheses (Cronbach, 1990). In this context, Nunnally and Bernstein (1994) argue that the development and validation of educational assessment techniques involve three critical phases: (i) to develop a specification table based on a conceptualization of an attribute based on substantive theory, (ii) to define the attribute in observable terms (e.g., questionnaire items), and (iii) to collect empirical data to verify that the measured attribute behaves in concordance with the underlying theory. Accordingly, the next section of the paper is concerned with the theoretical principles underlying the design of two questionnaires measuring student views of teacher behavior; it is an attempt not only to illustrate the process that was followed in order to develop the student questionnaires but also to reveal the two research areas upon which the design of questionnaires measuring student views on teacher behavior could be based. Specifically, drawing from the major findings of teacher effectiveness research a student questionnaire was developed. Moreover, a Greek version of the QTI (Wubbels & Levy, 1991) was developed to measure student views of their teacher interpersonal behavior. Thus, the next two sections of this paper provide a brief review of the literature on teacher effectiveness and teacher interpersonal behavior; and help us justify the theoretical framework upon which the design of each questionnaire was based. As far as the third phase is concerned, in order to ascertain the meaning of the scores generated from each questionnaire, analyses of their internal factor structures using structural equation modeling (SEM) procedures were undertaken.

Research Aims

In this context, the purpose of the study reported here was to refine a procedure that can be used to develop instruments for measuring student views of the quality of teacher behavior in the classroom and to examine their validity and reliability. It is also examined whether teacher effectiveness research or research into teacher interpersonal behavior can constitute bases for developing questionnaires to measure students' opinions about the behavior of their teachers in the classroom that can be of any use for establishing a valid teacher evaluation system. Thus, the various forms of the validity of each questionnaire are examined. In addition, this paper investigates the extent to which data on student views of teacher behavior collected from these two questionnaires are related to student achievement gains in cognitive and affective outcomes.

RESEARCH INTO TEACHER EFFECTIVENESS: FACTORS ASSOCIATED WITH TEACHER BEHAVIOURS

Information regarding the major findings of studies investigating the characteristics of effective teachers, conducted during the various phases of teacher effectiveness research, is provided here. The first stage of teacher effectiveness research was concerned with teachers' personal traits and led to presage-product studies, which attempted to identify the psychological characteristics of an effective teacher: personality characteristics (e.g., permissiveness, dogmatism, directness); attitude (e.g., motivation to teach, empathy toward children, and commitment); experience (e.g., years of teaching experience in grade level taught); and aptitude/achievement (e.g., professional recommendations, student teaching evaluations). Although this approach produced some consensus on virtues considered desirable in teachers, no information on the relations between these psychological factors and student performance was provided (Borich, 1992; Rosenshine & Furst 1973).

The subsequent focus produced experimental studies attempting to investigate the impact of specific teaching methods upon student achievement. However, the majority of these studies produced inconclusive results because the differences between teaching methods were not significant enough to produce meaningful differences in student achievement (Medley, 1979). Furthermore, the significant differences that did appear tended to contradict one another (Borich, 1992). The 1950s and 1960s brought concern about

creating a good classroom climate and about the teaching competencies involved in producing student achievement. This led to an emphasis on measurement of teacher behavior through systematic observation, and by 1970 to a proliferation of classroom observation systems (Shavelson, 1973; Simon & Boyer, 1970).

Research on "Process-Product": Factors Associated with Teacher Behaviours in Classroom

The last three decades researchers have turned to teacher behaviors as predictors of student achievement in order to build up a knowledge base on effective teaching. This research has led to the identification of a range of behaviors that are positively related to student achievement (Doyle, 1986; Brophy & Good, 1986; Everston , Anderson, Anderson, & Brophy, 1980; Borich, 1992; Galton, 1987). Many of these findings have been validated experimentally, but experimental findings are weaker and less consistent than correlational findings (e.g., Griffin & Barnes, 1986).

Quantity of Academic Activity

Brophy & Good (1986) argue that the most consistently replicated findings in American studies link student achievement to the *quantity and pacing of instruction*. Amount learned is related to opportunity to learn and achievement is maximized when teachers prioritize academic instruction (Brophy & Everston, 1976). Consistent success is another significant factor associated with student achievement. To learn efficiently, students must be engaged in activities that are appropriate in difficulty level and suited to their current achievement levels and needs (Bennett, Desforges, Cockburn, & Wilkinson, 1981; Stalling, 1985). Thus, there is a tension between the goal of maximizing amount of curriculum covered by pacing the students through the curriculum as rapidly as possible, and the need to move in small steps so that each new objective can be learned readily and without frustration. Brophy & Good (1986) argue that the pace at which a class can move should depend on the students' abilities and developmental levels and the nature of the subject matter because students' errors should be held to a minimum.

Classroom Management

Effective teachers are also expected to organize and manage the classroom environment as an efficient learning environment and thereby to maximize engagement rates (Creemers & Reezigt, 1996). Doyle (1986) points out that key indicators of effective classroom management include: good preparation of the classroom and installation of rules and procedures at the beginning of year; smoothness and momentum in lesson pacing; consistent accountability procedures; clarity about when and how students can get help; and about what options are available when they finish. As far as the actual teaching process is concerned, research into classroom discourse reveals that although in the classes of effective teachers there is a great deal of teacher talk most of it is academic rather than managerial or procedural, and much of it involves asking questions and giving feedback rather than extended lecturing (Cazden, 1986).

Quality of Teacher's Organized Lessons

The findings summarized above deal with factors associated with the quantity of academic activity. The variables presented below concern the form and quality of teacher's organized lessons, and can be divided into those that involve *giving information* (structuring), *asking questions* (soliciting), *providing feedback (*reacting), and *providing practice and application opportunities*. As for structuring, Rosenshine & Stevens (1986) point out that achievement is maximized when teachers not only actively present materials but structure it by: a) beginning with overviews and/or review of objectives; b) outlining the content to be covered and signaling transitions between lesson parts; c) calling attention to main ideas; and d) reviewing main ideas at the end. Summary reviews are also important because they integrate and reinforce the learning of major points (Brophy & Good, 1986). It can be claimed that these structuring elements not only facilitate memorizing information but allow for its apprehension as an integrated whole with recognition of the relationships between parts. Moreover, achievement is higher when information is presented with a degree of redundancy, particularly in the form of repeating and reviewing general views and key concepts. Clarity of presentation is also a consistent correlate of student achievement (Borich, 1992). Specifically, effective teachers are able to communicate clearly and directly to their students without digression, speaking above students' levels of comprehension or using speech patterns that impair the clarity of what is being taught (Smith & Land, 1981; Walberg, 1986).

Effective teachers are also expected to ask a lot of questions and to involve students in class discussion. Although the data on cognitive level of questions yield inconsistent results (Redfield & Rousseau, 1981), optimal question difficulty is expected to vary with context. There should also be a mix of product questions (i.e., expecting a single response from students) and process questions (i.e., expecting students to provide explanations) but effective

teachers are also expected to ask more process questions (Everston et al., 1980; Askew & William, 1995). Clarity of question is also a factor, and length of pause following questions should vary directly with their difficulty level. Specifically, a question calling for application of abstract principles should require a longer pause than a factual question (Brophy & Good, 1986).

Once the teacher has asked a question and called on a student to answer, the teacher has to monitor the student's response and react to it. Correct responses should be acknowledged for other students' learning, while responses that are partly correct, require affirmation of the correct part, and rephrasing of the question (Brophy & Good, 1986; Rosenshine & Stevens, 1986). Following incorrect answers, teachers should begin by indicating that the response is not correct but avoid personal criticism and show why the correct answer is correct (Rosenshine, 1971). In general, effective teachers are expected to answer relevant student questions or redirect them to the class and incorporate relevant student comments into the lesson (Brophy & Good, 1986; Borich, 1992; Flanders, 1970).

Effective teachers also use seatwork or small group tasks since they provide needed practice and application opportunities (Borich, 1992). The effectiveness of seatwork assignments are enhanced when the teacher explains the work that students are expected to do and once the students are released to work independently she/he circulates to monitor progress and provide help and feedback (Brophy & Good, 1986).

Classroom Climate

Classroom climate is a factor that teacher effectiveness research has found to be significant. The classroom environment should not be only *businesslike* but also needs to be *supportive* for the students (Walberg, 1986). Effective teachers expect all students to be able to succeed and their *positive expectations* should be transmitted to students. Finally, teachers are expected to *establish positive relationships with students* (Scheerens & Bosker, 1997). It can, therefore, be claimed that effective teachers are able to create a positive, learning-centered environment with an atmosphere of mutual respect between students and between students and the teacher (Kyriakides, Campbell, & Gagatsis, 2000). As a consequence, the environment is both safe and caring and the students feel that they are treated fairly by their teachers.

Beyond Classroom Behaviour

Factors other than classroom behavior have been the focus of considerable research effort. Although these factors can be classified in a variety of ways, the category system adopted here follows that used by Wang et al. (1990) who evaluated 179 authoritative papers examining the factors associated with student learning. The papers encompassed 228 items organized into 30 scales within six categories. Four of the categories related to beyond classroom factors and are concerned with two types of professional knowledge (i.e., subject knowledge and teacher's general knowledge of pedagogy), teachers' beliefs, and teachers' sense of efficacy.

Although teacher knowledge is widely perceived as a factor affecting teacher effectiveness (Scriven, 1994), the evidence for the effect of subject and/or pedagogy knowledge on student achievement is problematic. Moreover, an increasing amount of research into teacher effectiveness is focused on the beliefs of teachers (Askew et al., 1997). It is argued that teachers' own beliefs about and attitudes to teaching and the subject they teach are more important than immediately observable behaviors. The relationship between teachers' beliefs and practice is considered a dynamic two-way relationship because beliefs are also influenced by practical experience (Thompson, 1992; Rose et al., 1996). However, as Schoenfeld (1992) argued, the area of beliefs is under-conceptualized and needs new methodological and explanatory frames. In this context, teachers' sense of efficacy has recently become focus for research. Bandura (1997) defined perceived self-efficacy as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainment" (p. 3). In the same sense, teaching efficacy can be defined as teachers' beliefs in their capabilities to organize and orchestrate effective teaching–learning environments. Soodak and Podell (1996) argued that teacher efficacy influences several aspects of behavior that are important to teaching and learning. For example, it was shown that students with teachers who score high on self-efficacy did better on standardized tests of achievement (Anderson et al., 1988; Dempo & Gibson, 1985). Moreover, low teacher efficacy beliefs have been linked to low expectations of students, which is a significant factor predicting student achievement. However, further research is needed to identify the extent to which teacher self-efficacy is related to student achievement. As a consequence, this paper is only concerned with measuring student views of their teacher's behavior in classroom.

Research into Interpersonal Teacher Behaviour

The effect of teacher behavior on student outcomes has also been studied within the domain of classroom environment research, which found its origin in early teacher effectiveness studies and studies on the interaction between persons and environment (Moos, 1979; Walberg, 1979). Over the past 30 years, classroom environment research has shown the quality of the classroom environment in schools to be a significant determinant of student learning (Fraser, 1994; Dorman, 2003). For example, Goh and Fraser (1998) used the QTI to establish associations between student outcomes and perceived patterns of teacher-student interaction in primary school mathematics classes in Singapore. Thus, a particular line of research has evolved around order and classroom atmosphere, studying teaching in terms of the interpersonal relationship between teacher and students (Wubbels & Brekelmans, 1998; Wubbels & Levy, 1991). Specifically, in line with the systems approach to communication (Watzlawick, Beavin & Jackson, 1967) classroom groups are considered as ongoing systems. In the systems approach to communication the focus is on the effect of communication on the persons involved.

Therefore, to be able to describe the perceptions students have of the behavior of their teacher, Wubbels, Creton and Hooymayers (1987) applied a general model for interpersonal relationships designed by Leary (1957) to the context of education. In the Leary model, two dimensions are important. Leary called them the Dominance-Submission axis and the Hostility-Affection axis. While the two dimensions have occasionally been given other names, they have generally been accepted as universal descriptors of human interaction. Adapting the Leary Model to the context of education, Wubbels et al. (1987) used the two dimensions, which they called *Influence* (describing who is in control in the teacher-student relationship) and *Proximity* (describing the degree of cooperation between teacher and students). The influence dimension is characterized by teacher dominance (D) on one end of the spectrum, and teacher submission (S) on the other end. Similarly, the proximity dimension is characterized by teacher cooperation (C) on one end, and by teacher opposition (O) on the other. The two dimensions can be depicted in a two-dimensional plane, that can be further subdivided into eight categories or sectors of behavior: leadership (DC), helpful/friendly behavior (CD), understanding behavior (CS), giving responsibility/freedom (SC), uncertain behavior (SO), dissatisfied behavior (OS), admonishing behavior (OD) and strictness (DO). The Model for Interpersonal Teacher Behavior (MITB) also assumes that the eight sectors of behavior can be represented by two independent dimensions (i.e., Influence and Proximity). These eight categories of behavior are also expected to be ordered with equal distances to each other on a circular structure and maintain equal distances to the middle of the circle.

The QTI can be used to map students' perceptions of teacher interpersonal behavior according to the MITB. The QTI originally consisted of 77 items and a Likert-type 5-point scale is used to measure student views about these items. The items of the QTI refer to the eight sectors of behaviour, mentioned above, that jointly make up the MITB. Since its development, the QTI has been translated into more than 15 languages (Wubbels, Brekelmans, van Tartwijk, & Admiraal, 1997). Classroom environment and educational effectiveness studies that have included interpersonal teacher behaviour measured through the QTI have identified positive relationships between student perceptions of Influence and Proximity or their related (sub)sectors and cognitive student outcomes. For example, Brekelmans, Wubbels and Creton (1990) found that students' perceptions of teacher Influence were related to cognitive outcomes. Other, studies found positive correlations or regression coefficients for the leadership sector and cognitive student outcomes. Similar relationships have also been found for the Proximity dimension and Proximity related sectors such as helpful/friendly and understanding. Moreover, studies investigating the association of teacher-student relationship with affective outcomes reveal a much more consistent pattern than studies examining the effect of teacher-student relationships upon cognitive outcomes. However, the effects of teacher student relationships upon student outcomes found in the above studies were probably overestimated, because the nested structure of the data was not taken into account and multilevel analysis techniques were not used (Goldstein, 2003). In this context, this study attempts to identify the extent to which teacher effectiveness research and research into teacher interpersonal behavior can constitute bases for developing student questionnaires that can be used for conducting teacher evaluation. The extent to which student views of teacher interpersonal behavior is related to student achievement gains in both cognitive and affective outcomes is also examined.

RESEARCH DESIGN

The research reported here was conducted in five main stages.

Stage 1: A questionnaire measuring student views of teacher behavior based on the main findings of the process-product model of teacher effectiveness research was

constructed. Specifically, the items of the questionnaire covered: a) the quantity of academic activity; b) the form and quality of teacher's organized lessons; and c) classroom climate. Key indicators of the quantity of academic activity included: *quantity of instruction and smoothness and momentum in lesson pacing* and *classroom management*. Specifically, classroom management was measured by attempting to identify the extent to which teachers managed to establish consistent accountability procedures for maintaining attention on lesson and appropriate classroom behavior. Items concerning the form and quality of teacher's organized lessons were divided into those that involve teacher's skills in *giving information (structuring), asking questions (soliciting), providing feedback (reacting), and providing practice and application opportunities.* As far as the measurement of classroom climate is concerned, students were asked to provide information regarding the extent to which: a) the *classroom environment was businesslike and supportive* for the students; b) their teacher managed to establish *positive relationships with the students;* and c) *their teacher expects all students to be able to succeed*. Although it was not practical to include in the questionnaire items reflecting all the elements of quality of teaching, as emerged from teacher effectiveness research (see Scheerens & Bosker, 1997, pp. 123-133), it can be claimed that the nine indicators of the quality of teaching that were examined covered the most consistently replicated findings of TER presented above.

Stage 2: The content validity of the questionnaire measuring quality of teaching was determined by asking three researchers, two senior lecturers in pedagogy, two post-graduate students in education, and two primary inspectors, who were selected on the basis of their familiarity with the literature on teacher effectiveness, to evaluate the instrument in relation to two criteria. The validation specifications were: (1) each item should contain a recognizable generic teaching skill that could be easily observed from year-6 students; and (2) each item should contain one or more phrases that directly reflect a student's attitudes toward the way his/her teacher behaves in the classroom. The "judges" of the content validity of the questionnaire were asked to mark-up, make marginal notes, comments on or even rewrite the items. In the light of their comments, minor amendments were made, particularly where the structure used was not easily comprehensible, or terms that had been used were seen as not familiar to primary students. The final version of the questionnaire met the two criteria to the satisfaction of each of the nine "judges."

Stage 3: The development of a Greek-language version of the American 64 items-version of the QTI (Wubbels & Levy, 1991) followed procedures similar to those used by other researchers (Aldridge, Fraser, & Huang, 1998; Schibeci, Rideng, & Fraser, 1987) when they translated learning environment questionnaires. The first step was to check the relevance of each individual item of QTI and the whole concept (content validity) of each scale within the cultural context. With cross-cultural conceptual equivalence checked, items of the QTI were translated into Greek by an adult. The translator, fluent in both English and Greek, was a primary school teacher aware of many of the language difficulties that Cypriot primary students encounter. Many discussions between the researcher and the translator were held throughout this phase to ensure that the rewording of some items did not change their original meanings. A panel of judges, consisting of three primary teachers and two teacher educators who were fluent speakers of both languages, checked the preliminary translated version of QTI according to item clarity and face validity to refine the instrument, without sacrificing accuracy. The panel recommended further modification of some items. Back-translation procedures were then used. Back-translation is one of the few means of identifying when even standard words have different meanings for different groups (Smith, 1991). Two teacher educators who had not read the original English version of QTI translated the final translation of QTI from Greek back into English independently. The back-translation did not suggest that rewording was needed for any item.

Stage 4: Stratified sampling was used to select 38 Cypriot primary schools. Specifically, the Cypriot primary schools were divided into groups according to the location of the school (rural or urban) and the choice of the school sample in each group was random. All the year 6 students (N=1973) from each class (N=92) of the school sample were chosen. The chi-square test did not reveal any statistically significant difference between the research sample and the population in terms of students' sex (X^2=1.12, d.f.=1, p<.34). Moreover, the t-test did not reveal any statistically significant difference between the research sample and the population in terms of the age of students (t=0.27, d.f.=18925, p<.79). It may therefore be claimed that a nationally representative sample of Cypriot year-6 students was drawn. In March 2002, the final version of the questionnaire measuring student views about the quality of teaching was administered to the student of our sample. Specifically, Cypriot students were asked to indicate the extent to which his/her teacher behaves in certain ways when he/she teaches mathematics and Greek language on a 5-point scale (never=1, and 5=always). In May 2002, the students of our sample were also asked to complete the Greek version of the QTI to identify their interpersonal relationships with their mathematics teacher and their Greek language teacher. It is important to note that the students were asked to complete the questionnaires when

they were at the school (with permission given by the Ministry of Education) so that we had full data from 38 schools, 92 classes, and 1973 students.

Stage 5: The great majority of our sample (i.e., 32 schools out of 38 schools) participated in a study attempting to test the validity of the comprehensive model of educational effectiveness (Creemers, 1994) in relation to different criteria of measuring effectiveness (both cognitive and affective) (Kyriakides, 2005). Specifically, data on students' cognitive achievement in mathematics and Greek language were collected by using two forms of assessment (external assessment and teacher assessment), which were administered to them at the beginning and at the end of school year 2001-2002. Affective outcomes of schooling were measured through asking students to answer a questionnaire concerning their attitudes toward peers, teachers, school, and learning. It was therefore possible to examine the extent to which the answers of 1721 students in each of the two questionnaires mentioned above was associated with the effectiveness of their teachers in relation to different outcomes of schooling.

RESULTS

Results concerning the internal reliability and the discriminate and construct validity of the questionnaire used to measure student views of the quality of teaching are presented in the first part of the results section. The second part refers to the reliability and validity of the Greek version of QTI. Finally, the last part of this section is an attempt to identify the extent to which data collected from the two questionnaires are associated with student achievement gains in cognitive and affective outcomes.

The Questionnaire Measuring Student Views about the Quality of Teaching

Reliability, Consistency and Variance at Class Level

Although data collected from student responses to the questionnaire on quality of teaching are aggregated at the teacher/classroom level, scale internal consistency was calculated at individual student level for each subject. As a consequence to examine the reliability of the questionnaire, coefficient alpha values for the whole scale of the questionnaire and its subdomains for each subject were calculated. The values of Cronbach Alpha for the scale of the questionnaire for both subjects were very high (see Table 1). Moreover, the value of Cronbach Alpha for each subdomain was higher than 0.75 and can also be considered as satisfactory (Cronbach, 1990). It was also found that dropping any item from the overall scale of the questionnaire was not followed by a considerable increase in alpha value for any questionnaire or for any of their relevant subdomains.

A Generalizability Study on the use of the student questionnaire revealed that the data collected could be used for measuring the quality of teaching of each teacher in each

TABLE 1

Internal reliability for each scale of the questionnaire measuring the quality of teaching in each subject and for its total scale.

Scales	Cronbach Alpha	
	Mathematics	Language
Total scale score: Quality of teaching	.92	.91
Quantity and Pacing of Instruction (QPI)	.83	.82
Classroom Management (CM)	.82	.79
Giving Information (GI)	.79	.82
Asking Questions (AQ)	.82	.79
Providing Feedback (PF)	.85	.77
Providing Practice and Application Opportunities (PAO)	.86	.80
Creating a Businesslike and Supportive Environment (BSE)	.81	.85
Establishing Positive Relationships with Pupils (PRP)	.77	.80
Having Positive Expectations from Students (PES)	.78	.80

subject separately (Kyriakides, 2005). Thus, the score for each teacher in each questionnaire item was the mean score of the year-6 students of the class she/he taught.

Discriminate Validity

The mean correlation of one scale with the other scales measuring a multidimensional construct indicates the degree of discriminate validity. The lower the scales correlate amongst each other, the less they measure the same dimension of the construct. Thus, the discriminate validity was calculated for the nine student scales (see Table 2). We can observe that the student scales correlated between 0.10 and 0.45. Moreover, in each analysis, only 9 out of 36 correlations were statistically significant and all of them refer to the relationships of indicators of the same major-dimension of the quality of teaching (i.e., the quantity of academic activity, the form and quality of teacher's organized lessons, and classroom climate). Furthermore, the correlation coefficients that refer to the relationships of indicators of the same dimension of quality of teaching were higher than those that refer to the relationships of indicators of two different dimensions of quality of teaching. Finally, the values of the mean correlation of a scale with the other scales were smaller than .30. This implies that the 9 scales of the questionnaire, which refer to indicators of quality of teaching in each subject, differed sufficiently, although they partly measured the same general construct (i.e., the quality of teaching).

Construct Validity

Using a unified approach to test validation (AERA, APA and NCME, 1999; Messick, 1989), this study provides construct related evidence of the questionnaire measuring student views of the quality of teaching. For the identification of the factor structure of the questionnaire, SEM analyses were conducted using EQS (Bentler, 1995). Each model was estimated by using normal theory maximum likelihood methods (ML). The ML estimation procedure was chosen because it does not require an excessively large sample size. More than one fit index was used to evaluate the extent to which the data fit the models tested. Specifically, the scaled chi-square, Bentler's (1990) Comparative Fit Index (CFI), and the Root Mean Square Error of Approximation (RMSEA) (Brown & Mels, 1990) were examined. Finally, the factor parameter estimates for the models with acceptable fit were examined to help interpret the models.

Having in mind that analyses of structural equation models based on multiple scales provide more stable parameter estimates than models based on individual items (Rigdon, 1998), exploratory factor analyses of students' responses to the 76 items of the questionnaire measuring the quality of teaching were conducted. For each subject, based on the results of the factor analysis (explaining 78% of the total variance in the case of Mathematics and 82% in the case of Greek Language), three mean scores representing students' views on each of the nine theoretically postulated subdomains of the inventory shown in Table 1 were created. Thus, a first-order Confirmatory Factor Analysis model designed to test the multidimensionality of the questionnaire (Byrne, 1998) was used to examine the construct validity of the questionnaire. Specifically, the model hypothesized that: (a) the 27 variables (i.e., mean scores) could be explained by nine factors concerning the nine subdomains of the questionnaire used to measure quality of teaching; (b) each variable would have a nonzero loading on the factor it was designed to measure, and zero loadings on all other factors; (c) the nine factors would be correlated; and (d) measurement errors would be uncorrelated.

The findings of the first order factor SEM analysis generally affirmed the theory upon which the questionnaire was developed. Although the scaled chi-square for the nine-factor structure in each subject (Mathematics: X^2=508.8, d.f.=288, p<.001; Greek Language: X^2=495.3, d.f.=288, p<.001) as expected was statistically significant, the values of RMSEA (Mathematics: 0.031 and Greek Language: 0.029) and CFI (Mathematics: 0.979 and Greek Language: 0.981) met the criteria for acceptable level of fit. Kline (1998, p. 212) argues that "even when the theory is precise about the number of factors of a first-order model, the researcher should determine whether the fit of a simpler, one-factor model is comparable." Criteria fit for a one-factor model (Mathematics: X^2=1549.4, d.f.=324, p<.001; RMSEA=0.144 and CFI=0.455; Greek Language: X^2=1364.8, d.f.=324, p<.001; RMSEA=0.142 and CFI=0.435) provided values that fell outside generally accepted guidelines for model fit. Thus, a decision was made to consider the nine-factor structure as reasonable in both cases and the analysis proceeded and the parameter estimates were calculated. Figures 1 and 2 depict the nine-factors model and present the factor parameters estimates for each subject separately. All parameter estimates were statistically significant (p<.001).

The following observations arise from figures 1 and 2. First, the standardized factor loadings were all positive and moderately high. Their standardized values ranged from 0.59 to 0.78 and the great majority of them were higher than 0.65. Second, the correlations among the nine factors were positive and ranged between 0.27 and 0.42. Moreover, the majority of factor inter-correlations were higher than 0.30.

TABLE 2

Correlation coefficients between the student scales and average correlation between the student scales in each subject .

Scale	Correlation Coefficients								Average
	2	3	4	5	6	7	8	9	correlation
A) Mathematics									
1	.36**	.21	.27	.25	.14	.26	.12	.19	.225
2		.11	.19	.24	.17	.26	.10	.13	.195
3			.42**	.33**	.38**	.29	.19	.21	.268
4				.34**	.24	.21	.16	.20	.254
5					.32**	.26	.21	.18	.266
6						.24	.18	.23	.238
7							.31**	.44**	.284
8								.39**	.208
9									.246
B) Greek Language									
1	.34**	.20	.25	.28	.16	.22	.10	.17	.215
2		.12	.18	.14	.22	.25	.11	.15	.189
3			.44**	.33**	.32**	.25	.24	.26	.271
4				.21	.37**	.18	.19	.22	.255
5					.33**	.28	.23	.18	.249
6						.26	.15	.25	.258
7							.36**	.40**	.275
8								.45**	.229
9									.260

Note:

A) Scales 1 up to 2 refer to the following indicators of quantity of academic activity:
1 = Quantity of instruction and smoothness and momentum in lesson pacing, and 2= Classroom Management.

B) Scales 3 up to 6 refer to the following indicators of the form and quality of teacher's organized lessons:
3 = Giving information, 4 = Asking questions, 5 = Providing feedback, 6 = Providing practice and application opportunities

C) Scales 7 up to 9 refer to the following indicators of classroom climate:
7 = Creating a businesslike and supportive environment, 8 = Establishing positive relationships with pupils and 9 = Having Positive expectations from students

** $p < .001$

FIGURE 1

First-order factor model of the questionnaire measuring quality of teaching in Mathematics with factor parameter estimates.

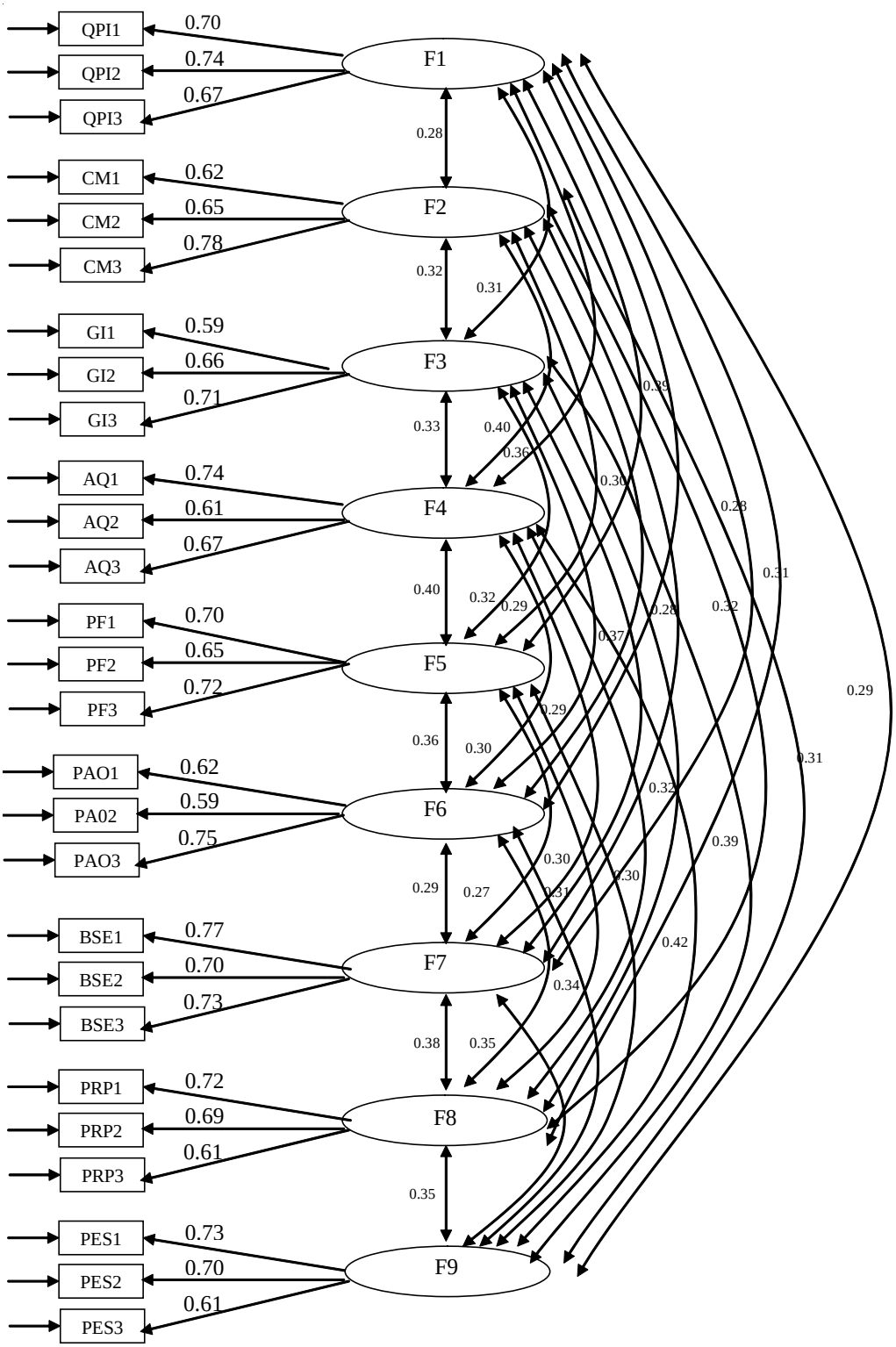
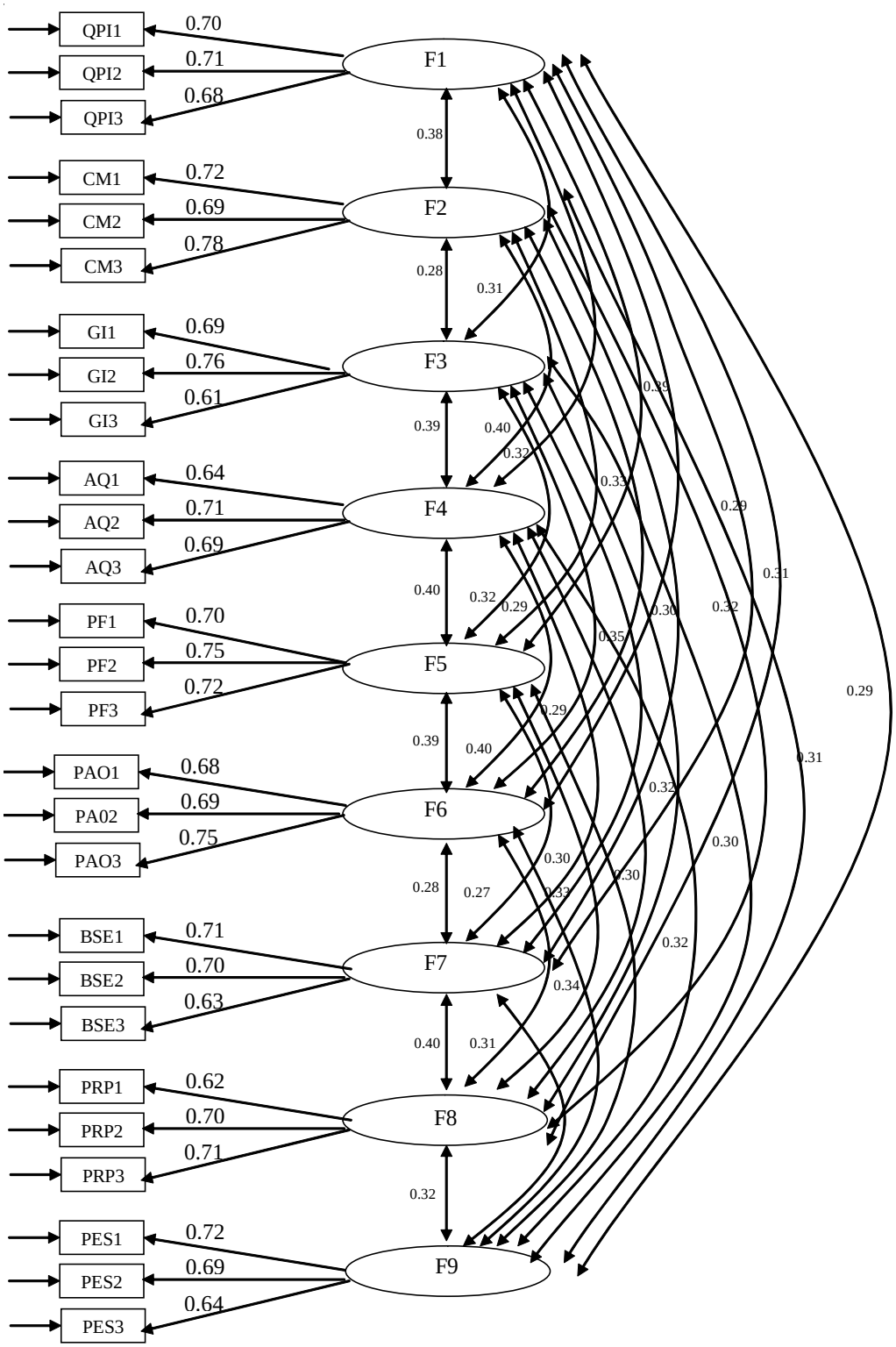


FIGURE 2

First-order factor model of the questionnaire measuring quality of teaching in Greek Language with factor parameter estimates.



It was therefore decided to examine whether the second-order factor model could explain the relatively high correlations among the nine first-order factors.

The relatively high values of the factor intercorrelations provided further support to our attempt to identify a higher order model, which could explain the correlations among the nine first-order factors in each analysis. This model hypothesized that for each subject: (a) responses to the student questionnaire could be explained by nine first-order factors and one second-order factor (i.e., quality of teaching in general); (b) each item (i.e., scale score) would have a nonzero loading on the factor it was designed to measure, and zero loadings on all other factors; (c) error terms associated with each item would be uncorrelated, and (d) covariation among the nine first-order factors would be explained by their regression on the second order factor.

Figures 3 and 4 illustrate the models with one second-order factor for each subject. The fit statistics for both analyses (Mathematics: scaled X^2=540.4, d.f.=315, p<.001; RMSEA=0.028 and CFI=0.982; Greek Language: scaled X^2=537.6, d.f.=315, p<.001; RMSEA=0.027 and CFI=0.985) were acceptable. By comparing the second-order factor model that emerged from analyzing the data of each subject with its theoretical first-order factor model, we could identify a minor decrease of the RMSEA (i.e., Mathematics: from 0.031 to 0.028 and Greek Language: from 0.029 to 0.027) and a very minor increase of the CFI (i.e., Mathematics: from 0.979 to 0.982 and Greek Language: from 0.981 to 0.985). Thus, the single second order model was considered as appropriate (Maruyama, 1998) and thereby the analysis proceeded and the parameter estimates were calculated.

The following observations arise from figures 3 and 4. In each model, the great majority of the standardized factor loadings were higher than 0.65. Moreover, the standardized factor loading revealed that the model explained more than 50% of variance of at least 15 items in each subject. Finally, the standardized path coefficients relating the first-order factors to the second-order factors were higher than 0.50. It is finally important to note that in terms of the theory upon which the questionnaire was based, the second-order factor, which has been identified through analyzing data emerged from each subject separately, represents the quality of teaching in each subject.

Sensitivity of Student Ratings

Sensitivity refers to whether students are able to discriminate between teachers. It is important to note that in order to analyze the sensitivity of student ratings, data collected through the last stage of the research design were analyzed. Specifically, by taking into account student prior knowledge in each subject, it was found that the differences between classes for each scale were statistically significant. Moreover, the differences between scales are the strongest in relation to scale 3 [Giving Information (Mathematics: F = 15.8, Greek Language: F = 14.2)], scale 5 [Providing Feedback (Mathematics: F = 16.1, Greek Language: F = 13.3)], scale 7 [Creating businesslike and supportive environment (Mathematics: F = 12.4, Greek Language: F = 14.5)], and the weakest in scale 2 [Classroom Management (Mathematics: F = 2.4, Greek Language: F = 2.2)].

The Greek Version of the QTI

Taking into account the main assumptions of the theoretical framework upon which the design of the QTI was based, the reliability and validity of the Greek version of the QTI can be examined. As it was mentioned above, the QTI can be used to measure students' perceptions of teacher interpersonal behavior according to the MITB, which is a special model because of its statistical properties. This model is theoretically linked to a particular branch of models called *circumplex models*. Thus, the design of QTI is based on four assumptions. First, the eight behavioral sectors (or scales) of the model are represented by two dimensions (or factors). Second, the two interpersonal dimensions that lay behind the sectors are uncorrelated. Third, the sectors (or scales) of the model can be ordered in a circular structure. Finally, the sectors (or scales) of the model are equally distributed over this circular structure. Thus, the investigation of the reliability and validity of the Greek version of QTI was based on an attempt to test these assumptions.

Reliability, Consistency and Variance at Class Level

A learning environment instrument is expected to differentiate between perceptions of students in different classes. Students within a class usually view the learning environment similarly but differently from students in other classes. Thus, reliability was computed for each of the scales of the QTI by calculating multilevel λ (Snijders & Bosker, 1999) and Cronbach alpha for data aggregated at the class level. The value of Cronbach alpha represents consistency across items whereas multilevel λ represents consistency across groups of students. The results are presented in Table 3. We can observe that reliability coefficients were very high (around .90). Moreover, the reliability of the scales Giving Responsibility/Freedom (SC) and Strict (DO) were somewhat lower, while the reliability of the scale Leadership (DC) was the highest.

FIGURE 3

Second-order factor model of the questionnaire measuring quality of teaching in Mathematics with factor parameter estimates.

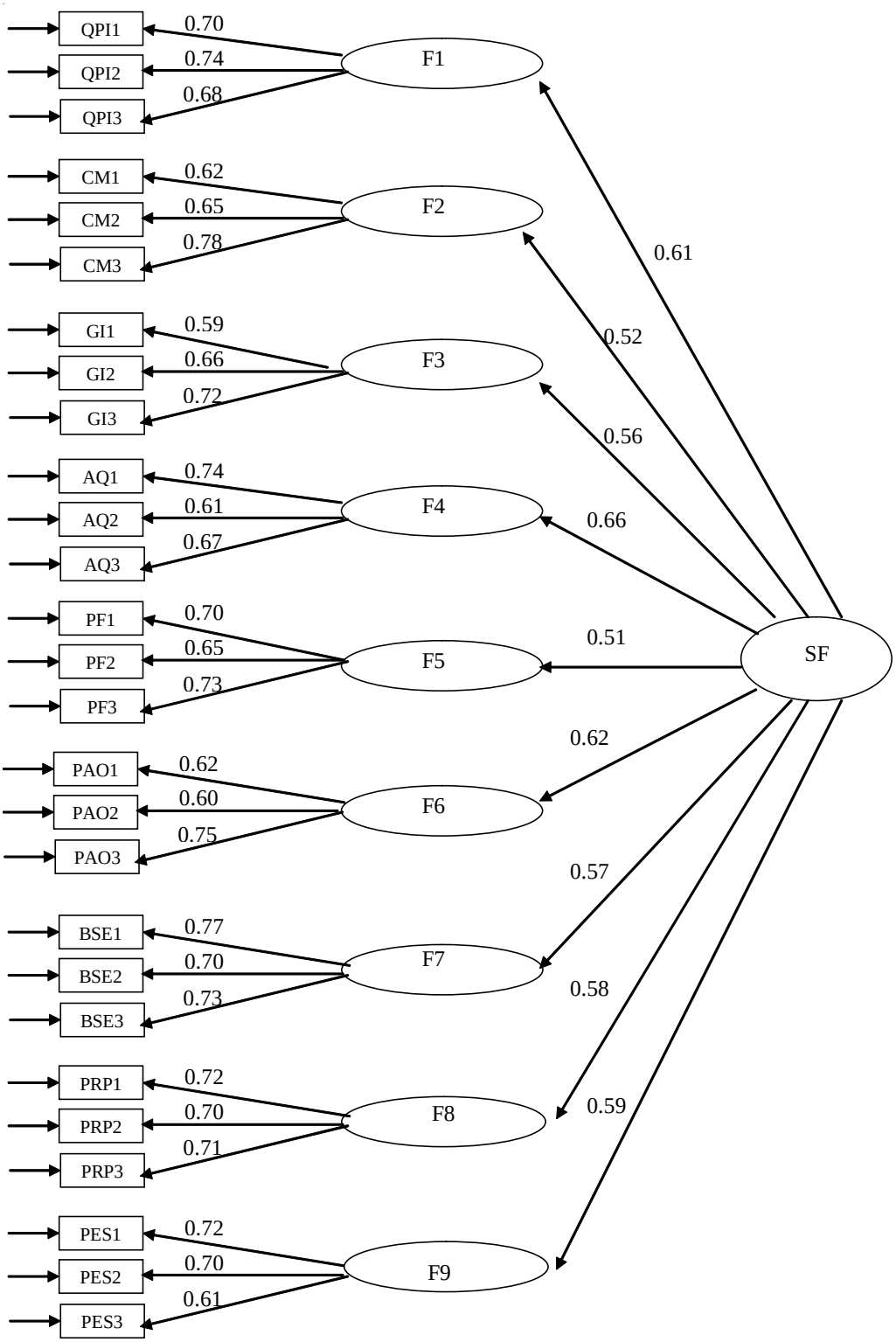


FIGURE 4

Second-order factor model of the questionnaire measuring quality of teaching in Greek Language with factor parameter estimates

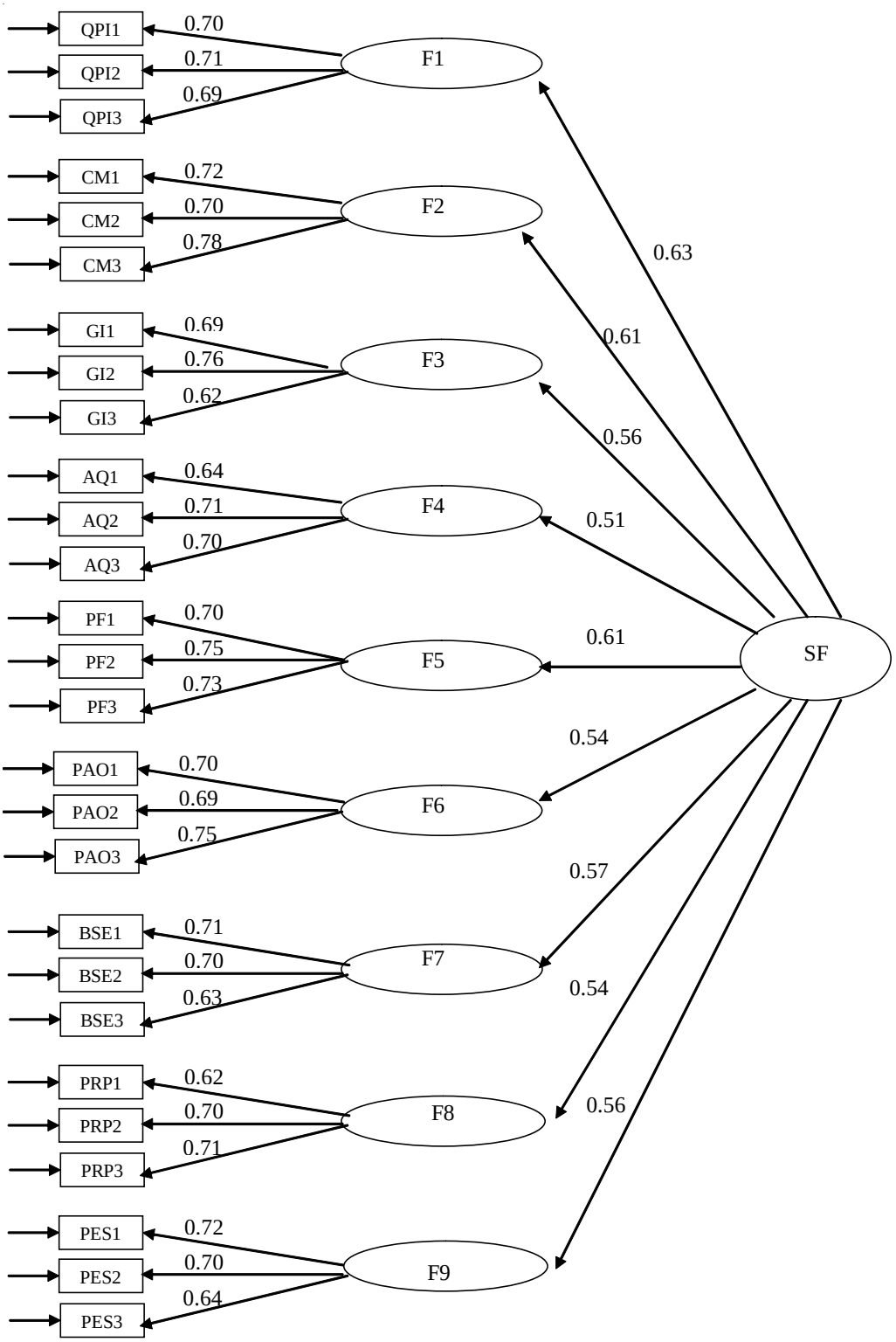


TABLE 3

Cronbach alpha (reliability), Multilevel Lambda (consistency), and intra-class correlations (ICC) of QTI scales at the teacher/class level in each subject.

Scales	Mathematics			Greek Language		
	Cronbach α	λ	ICC	Cronbach α	λ	ICC
DC	.94	.93	.46	.92	.93	.47
CD	.91	.91	.41	.90	.92	.40
CS	.92	.88	.42	.91	.90	.42
SC	.84	.85	.36	.85	.88	.36
SO	.90	.90	.43	.91	.92	.41
OS	.91	.89	.41	.93	.89	.39
OD	.90	.89	.39	.92	.90	.40
DO	.88	.87	.39	.87	.86	.36

Note: The scales of the QTI are as follows: leadership (DC), helpful/friendly behavior (CD), understanding behavior (CS), giving responsibility/freedom (SC), uncertain behavior (SO), dissatisfied behavior (OS), admonishing behavior (OD) and strictness (DO).

Using the Mplus (Muthén & Muthén, 1999) the intra-class correlations of the scales were computed. The intra-class correlations, which indicate what amount of variance of the QTI is located at the between level, are also illustrated in Table 3. We can observe that the percentages of variance at the between level (teacher-class level) were between 36 and 46 for the mathematics teachers' data and between 36 and 45 for the Greek Language teachers' data. These percentages are rather high compared to other instruments that measure perceptions of people or objects in clustered or interdependent situations.

Construct Validity

Construct validity of the QTI was investigated by subjecting the scale scores to a multilevel factor analysis using Mplus. From these analyses, it was found that an unequally-spaced circumplex model fitted the data well (Mathematics: $X^2 = 72.15$ d.f.= 13, $p<.001$; CFI=.988; RMSEA=.041; and Greek Language: $X^2 = 50.59$, d.f. = 13, $p<.001$, CFI=.991; RMSEA=.038). The factor loadings resulting from this model are also presented in Table 4. This model is based on the assumption that the eight scale scores are ordered in a circle and represented by two independent dimensions. It is however not found that the eight sectors of teacher interpersonal behavior are equally distributed over the circle or equally distanced to the circle center. This implies that student responses to the Greek version of QTI helped us generate empirical evidence supporting the first three assumptions upon which the design of QTI was based.

TABLE 4

Factor loadings for the unequally spaced Circumplex Model.

	Mathematics		Greek Language	
Scales	Factor 1	Factor 2	Factor 1	Factor 2
DC	1.00	.33	1.00	.51
CD	.29	0.96	.35	1.05
CS	.09	1.02	.02	1.00
SC	-.39	.53	-.44	.59
SO	-0.99	.14	-1.00	-.16
OS	-.18	-.73	-.08	-.70
OD	-.04	-.88	.04	-.78
DO	.34	-.65	.56	-.52

Note: The meaning of each QTI scale is shown in Table3.

Therefore, the two dimension scores, rather than the eight sector scores, can be used to evaluate teacher interpersonal behavior and to identify the effects of interpersonal teacher behavior on student achievement gains in each subject.

The Effect of Different Measures of Teacher Behaviour Upon Student Achievement Gains

Since the two questionnaires measuring student views about different aspects of teacher behavior were also used to collect data for the purposes of an effectiveness study (Kyriakides, 2005), this section is concerned with the extent to which data emerged from the two student questionnaires are associated with student achievement gains in cognitive and affective outcomes of schooling. To examine the extent to which variables measuring teacher behaviors show the expected effects upon student achievement, the analyses were performed separately for each dependent variable. Thus, the first step in the analysis was to determine the variance at individual, class and school level without explanatory variables (empty model). In each of the three empty models, the variances at each level reached statistical significance ($p<.001$). This implies that MLwiN can be used to identify the explanatory variables that are associated with student achievement in each outcome measure. Specifically, of the total variance in each outcome measure, the variance at school level was 11.5% in Mathematics, 9.7% in Greek Language and 14.3% in affective outcomes. The variance at classroom level was 15.2% in Mathematics, 16.8% in Greek Language and 17.8% in affective outcomes. This implies that in Cyprus the effect of both the school and the classroom was more pronounced on achievement in affective outcome measures rather than in cognitive measures in mathematics and Greek language.

In subsequent steps explanatory variables at different levels were added, starting at the student level. Explanatory variables, except grouping variables, were centered as Z-scores with a mean of 0 and a standard deviation of 1. This is a way of centering around the grand mean (Bryk & Raudenbush, 1992) and yields effects that are comparable. Thus, each effect expresses how much the dependent variable increases (or decreases in case of a negative sign) by each additional deviation on the independent variable (Snijders & Bosker, 1999). Grouping variables were entered as dummies with one of the groups as baseline (e.g., boys=0). It is important to note that various explanatory variables, which can be categorized as context, time, opportunity and quality factors, were taken into account to test the main assumptions of Creemers' model (see Kyriakides, 2005). However, only the effect of student background factors and the effects of the various measures of teacher behavior upon student achievement in Mathematics, Greek Language and affective outcomes of schooling are shown in Table 5.

In model 1 the context variables at student level (i.e., SES, prior knowledge, sex) were added to the empty model. Variables concerned with the context of each classroom, such as the average baseline score, the average SES score, and the percentage of girls were also added to the empty model. We can observe that all three contextual factors at student level (i.e., SES, prior knowledge, sex) had a significant effect upon achievement in each of the three outcome measures. Moreover, SES was correlated to a much higher degree with cognitive than affective achievement gains. Furthermore, the effect of gender background was not the same, since girls achieved lower scores than boys in mathematics and higher scores in Greek language and in affective aims of schooling. As far as the effect of the contextual factors at classroom level is concerned, only the average SES and the average baseline score were found to be associated with student achievement. Finally, model 1 helped us explain almost half of the total-variance of student achievement in each outcome measure and most of which was at the student level.

In model 2, explanatory variables concerned with the quality of teaching that emerged from the student questionnaire were entered. We can observe that eight out of the nine scales of the student questionnaire were associated with student achievement gains in both Mathematics and Greek Language. The scale measuring teachers' positive expectations from their students did not have any statistically significant effect upon student achievement in mathematics and Greek language. As far as the effect of these variables on the affective outcomes of schooling is concerned, the four scales, which refer to the form and quality of teacher's organized lessons did not have any statistically significant effect. Moreover, the effect of each of the three scales measuring classroom climate on student achievement in affective outcomes was stronger than the effect of the scales measuring the quantity of academic activity. Finally, this model helped us explain more than 5% of the total variance in each outcome measure and most of it was at the classroom level. However, in each analysis more than 6% of the total variance remained unexplained at the classroom level

TABLE 5

Parameter Estimates and (Standard Errors) for the models* used to investigate educational effectiveness in each outcome of schooling.

	Mathematics				Greek Language				Affective outcomes			
Factors	**Model 0**	**Model 1**	**Model 2**	**Model 3**	**Model 0**	**Model 1**	**Model 2**	**Model 3**	**Model 0**	**Model 1**	**Model 2**	**Model 3**
Fixed part (Intercept)	37.4 (1.3)	37.2 (1.2)	39.4 (1.0)	40.4 (0.8)	34.4 (1.3)	35.2 (1.0)	36.6 (.72)	37.2 (.61)	34.4 (1.3)	35.2 (1.0)	36.4 (.81)	36.8 (.71)
Student Level												
Context												
Baseline score		2.18 (.09)	2.10 (.08)	2.03 (.07)		2.08 (.09)	2.07 (.08)	2.07 (.08)		2.38 (.08)	2.39 (.09)	2.39 (.08)
Sex (Boys = 0, Girls = 1)		-0.81 (.17)	-0.78 (.07)	-0.76 (.07)		0.92 (.15)	0.88 (.15)	0.89 (.15)		0.90 (.18)	0.88 (.18)	0.89 (.18)
Socio Economical Status (SES)		2.03 (.12)	1.96 (.10)	1.96 (.10)		2.06 (.11)	2.06 (.11)	2.07 (.11)		1.26 (.14)	1.27 (.14)	1.27 (.13)
Classroom Level												
Context												
Average baseline score		2.30 (.39)	2.31 (.33)	2.30 (.32)		2.13 (.35)	2.12 (.34)	2.10 (.33)		2.04 (.33)	2.08 (.33)	2.10 (.33)
Average SES		1.43 (.44)	1.45 (.42)	1.46 (.42)		1.64 (.43)	1.63 (.43)	1.63 (.42)		1.05 (.41)	1.13 (.42)	1.43 (.42)
Percentage of girls		N.S.S.**	N.S.S.	N.S.S.		N.S.S.	N.S.S.	N.S.S.		N.S.S.	N.S.S.	N.S.S.
Quality of teaching												
Quantity and pacing of instruction			0.81 (.09)	0.81 (.09)			0.80 (.09)	0.81 (.09)			0.70 (.12)	0.71 (.12)
Classroom Management			0.90 (.09)	0.92 (.09)			0.94 (.10)	0.92 (.09)			0.73 (.11)	0.72 (.10)
Giving Information			0.96 (.08)	0.97 (.08)			0.97 (.08)	0.97 (.08)			N.S.S.	N.S.S.
Asking Questions			0.95 (.10)	0.95 (.09)			1.01 (.09)	1.07 (.09)			N.S.S.	N.S.S.
Providing Feedback			0.91 (.10)	0.92 (.10)			1.08 (.10)	1.08 (.09)			N.S.S.	N.S.S.
Practice and application			0.97 (.09)	0.99 (.09)			0.93 (.09)	0.93 (.09)			N.S.S.	N.S.S.
Classroom environment			0.84 (.10)	0.84 (.09)			0.82 (.09)	0.84 (.09)			0.94 (.09)	0.94 (.09)
Positive relationships with students			0.73 (.09)	0.73 (.09)			0.69 (.09)	0.69 (.08)			0.97 (.08)	0.99 (.08)
Positive expectations from students			N.S.S.	N.S.S.			N.S.S.	N.S.S.			0.98 (.08)	N.S.S.
Teacher Interpersonal Behavior												
DS (Influence)				0.81 (.10)				N.S.S.				0.99 (.08)
CO (Proximity)				N.S.S.				0.84 (.09)				0.77 (.08)
Variance components												
School	11.5%	7.8%	7.6%	7.5%	9.7%	8.9%	8.7%	8.7%	14.3%	13.1%	12.9%	12.9%
Class	15.2%	10.4%	6.3%	5.0%	16.8%	12.5%	6.1%	5.0%	17.8%	14.6%	11.1%	8.9%
Student	73.3%	30.3%	28.3%	28.0%	73.5%	27.6%	25.1%	24.3%	67.9%	24.0%	22.4%	21.2%
Absolute	134.41	65.18	56.72	54.44	154.23	75.57	61.54	58.61	140.29	72.53	65.10	60.32
Explained		51.5%	57.8%	59.5%		51.0%	60.1%	62.0%		48.3%	53.6%	57.0%
Significance test												
X^2	1225.60	800.65	691.05	678.73	1045.27	633.14	553.02	543.01	1124.25	812.12	721.00	710.68
Reduction		424.95	109.60	12.32		412.13	80.12	10.01		312.13	91.12	10.32
Degrees of freedom		5	8	1		5	8	1		5	5	1
p-value		.001	.001	.001		.001	.001	.001		.001	.001	.001

The models were estimated without the variables that did not have a statistically significant effect. ** N.S.S. = No statistically significant effect (i.e. p>.05)

In model 3, the variables measuring teacher influence and proximity, which emerged from student responses to the Greek version of QTI, were entered. Teacher influence was found to be associated with student achievement in mathematics and affective outcomes of schooling whereas proximity was associated with achievement in Greek language and affective outcomes of schooling. In the analysis of student achievement in affective outcomes of schooling, when "teacher proximity" was entered the effect of "positive expectations from students" was disappeared. We can also observe that the explanatory variables emerged from student responses to the QTI helped us explain 3.4% of the total variance in achievement of affective outcomes but less than 2% of the total variance in cognitive outcomes. However, in each analysis, the likelihood statistic reveals a statistically significant reduction ($p<.001$) from model 2 to model 3, which justifies the selection of model 3.

DISCUSSION

The evidence previously presented is discussed in terms of its implications for the development of a teacher evaluation system based on multiple data sources. The study reported here revealed that Cypriot students of year 6 are able to provide reliable and valid data on the behavior of their teachers, which can help us evaluate the quality of teaching and the interpersonal teacher behavior. Moreover, suggestions for establishing mechanisms for evaluating the quality of student ratings are provided. It is argued that rather than investigating the sensitivity of student ratings and student ability to provide reliable data about the behavior of their teacher, meta-evaluation mechanisms of student ratings should examine the content of the instruments used to collect data. It was also shown that teacher effectiveness research and research into teacher interpersonal behavior can help us build the theoretical background upon which the design of student questionnaires measuring teacher behavior can be based.

Specifically, a questionnaire based on the main findings of teacher effectiveness research was designed and empirical evidence supporting both the content and the construct validity of the questionnaire was provided. In addition, data emerged from student responses to the questionnaire were associated with student achievement gains in both cognitive and affective outcomes of schooling. This implies that teacher evaluation data emerged from this questionnaire are in line with data on value-added assessment of student achievement, which help us measure teacher effectiveness. Similarly, this study illustrates how the reliability and validity of the Greek version of QTI, which is based on the circumplex MITB, can be examined. Moreover, empirical evidence, supporting that reliable and valid data emerged from Cypriot student responses to the QTI, is provided. Thus, the QTI scales were found to be represented by two independent dimensions: influence and proximity. Finally, results of multilevel analyses revealed that influence and proximity were positively related to both cognitive and affective outcomes of schooling. This implies that value-added data of student achievement are associated with data emerged from student responses to the two questionnaires, which were designed by taking into account the major findings of two different theoretical domains: teacher effectiveness research and research into teacher interpersonal behavior.

It is important to acknowledge the various methodological limitations of using student performance data in teacher evaluation (Gray, Goldstein, & Thomas, 2001; Goldstein, 2001). These limitations partly arise from the technical and practical difficulties of measuring teacher effectiveness through multilevel modeling techniques that take into account those student, classroom, and school variables, which, according to the research findings, have significant effects on student achievement gains (Kyriakides & Campbell, 2003). The difficulties of collecting such information and of using them are also reflected to the fact that even in a relatively sophisticated evaluation system such as the system of the United Kingdom, Goldstein (2001) demonstrated naive and inappropriate use of such data by the government. In addition, Gray et al. (2001) showed the difficulty in assuming stability and consistency of effectiveness measurements to predict future performance. Moreover, one of the primary criticisms of using the value-added approach to evaluate teachers, concerns the fact that it is not easily explicable to those who are most affected by it (Baker et al., 1995). In Tennessee, there is an ongoing real life "experiment" that illustrates difficulties in the practical application of multilevel modeling to indicator systems: The Tennessee Value Added Assessment System (TVAAS) (Sanders and Horn, 1994). However, these difficulties are not merely about technical accessibility (Millman, 1997); controversy in the USA in using this approach for teacher evaluation also had political dimensions (Teddlie, Reynolds, & Sammons, 2000).

This study has shown that student ratings of teacher behavior are highly correlated with value-added measures of student cognitive and affective outcomes. This implies that student ratings rather than value-added measures of student outcomes can be considered as a more practical and

valid way of evaluating teachers. Nevertheless, a critical issue in using student ratings to measure teacher effectiveness seems to be the theoretical background upon which the design of student questionnaires can be based. It can be argued that educational effectiveness research and research into teacher interpersonal behavior could help us establish the theoretical framework upon which a more valid teacher evaluation system could be built.

However, an important constraint of the approach used in this study to evaluate the quality of data emerged from student ratings has to do with the fact that the impact of collecting such data on teacher professional development has not been examined. This implies that longitudinal studies should be conducted in order to examine whether feedback from students emerged from the questionnaires presented here could be used for instructional development of the teachers of our sample. Such data can help us identify the extent to which the questionnaire measuring quality of teaching and the QTI could serve the formative purposes of teacher evaluation. The importance of such research is also supported by the fact that there is some evidence showing that instructional improvement does not take place automatically as a result of giving questionnaires to students and presenting an instructor with a printed summary of results (Fresko & Nasser, 2001). It is therefore important to find out why feedback from student ratings can not be easily translated into improvement of teaching practice. Such studies can also help find ways to help teachers adequately understand how to read data emerging from student ratings. Teachers may use defense mechanisms, such as denial, repression, and rationalization, when faced with less than positive feedback from their students. It is always easier to adhere to familiar ways and to continue doing what one has always done as opposed to attempting something new with unknown consequences. Additional empirical evidence is therefore needed to determine how multiple measures of teacher performance and especially student ratings can contribute to the process of teacher professional development. Moreover, a study on Cypriot teachers' perceptions of criteria of teacher evaluation revealed that they did not consider the criteria reflecting students' satisfaction as appropriate for their evaluation (Kyriakides, Demetriou, & Charalambous, 2006). Thus, before attempting to introduce a new system of teacher evaluation based on multiple measures of teachers' performance, it will be necessary to persuade teachers that the use of student ratings of teacher behavior can provide them with additional feedback and assessment information, both for personal and professional improvement and for ensuring accountability in performance (Peterson et al., 2000).

Finally, implications of findings for the development of educational effectiveness research can be drawn. This study has shown that data on teacher interpersonal behavior emerged from student responses to the Greek version of QTI helped explain variance on student achievement in both cognitive and affective outcomes of schooling. This implies that interpersonal behavior as perceived by students may be an important variable for educational effectiveness researchers. However, national and comparative studies should be conducted to identify the importance of treating variables associated with teacher interpersonal behavior as educational effectiveness factors. Research from cross-national studies and cross-cultural studies using the QTI indicate that the instrument and model are cross-culturally valid (den Brok, Levy, Wubbels, & Rodriguez, 2003; Wubbels & Levy, 1991). This implies that researchers can use the QTI in large-scale international effectiveness studies that may help establish the international dimension of educational effectiveness research. Such studies may also help find the extent to which teacher interpersonal behavior explains effectiveness across countries, and whether it should be included in generic or differentiated models of educational effectiveness.

REFERENCES

Aldridge, J. M., Fraser, B. J., & Huang, T. I. (1998). *A cross-national study of perceived classroom environments in Taiwan and Australia*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Aleamoni, L.M. (1981). Student rating of instruction. In J. Millman (Ed) *Handbook of teacher evaluation, 110-145*. London: Sage.

Aleamoni, L.M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*(2), 153 – 166.

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory.* Monterey: Brooks and Cole

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anderson, R., Green, M., & Loewen, P. (1988). Relationships among Teachers and Students Thinking skills, Sense of Efficacy and Student Achievement. *Alberta Journal of Educational Research, 17*, 86-95.

Askew, M. & William, D. (1995). *Recent Research in Mathematics Education* 5-16. London: Office for Standards in Education.

Askew, M., Rhodes, V., Brown, M., William, D., & Johnson, D. (1997). *Effective teachers of numeracy: Report of a study carried out for the teacher training agency.* London: Kings College London School of Education.

Baker, A. P., Xu, Dengke, & Detch, E. (1995). *The measure of Education: A review of the Tennessee Value Added Assessment System.* Office of Education Accountability, Department of Education, Nashville, TN: Comptroller of the Treasury.

Bandura, A. (1997). *Teacher's sense of Efficacy: An important factor in school Achievement.* New York: W.H. Freeman and Company.

Bennett, N., Desforges, C., Cockburn, A. & Wilkinson, B. (1981). *The quality of pupil learning experience: Interim Report.* Lancaster: University of Lancaster, Centre for Educational Research and Development.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M. (1995). *EQS: Structural equations program manual.* California: Multivariate Software Inc.

Borich, G. D. (1992) (2nd Ed). *Effective teaching methods.* New York: Macmillan Publishing Company.

Brekelmans, M., Wubbels, Th. & Créton, H.A. (1990). A study of student perceptions of physics teacher behavior. *Journal of Research in Science Teaching, 27*, 335-350.

Brophy, J. & Everston, L. (1976). *Learning from Teaching: A Developmental Perspective.* Boston: Allyn and Bacon.

Brophy, J. & Good, T. L. (1986). Teacher Behavior and Student Achievement. In M.C. Wittrock (Ed.) *Handbook of Research on Teaching* (pp. 328-375). New York: MacMillan.

Brown, M. W., & Mels, G. (1990). *RAMONA PC: User Manual.* Pretoria: University of South Africa.

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models.* Newbury Park: CL: SAGE.

Byrne, B. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications and programming.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cazden, C. B. (1986). Classroom Discourse. In M. C. Wittrock (Ed.) *Handbook of Research on Teaching* (pp. 432-463). New York: MacMillan.

CEPI (2000). *Teacher Evaluation*. [Online]. Available: www.edpolicyvcu.org/policy_issues/ staffing/ p_teacher_eval.html (January 27, 2003).

Creemers, B.P.M. (1994). *The effective classroom*. London: Cassell.

Creemers, B.P.M. & Reezigt, G.J. (1996). School level conditions affecting the effectiveness of instruction. *School Effectiveness and School Improvement*, 7, 197-228.

Cronbach, L.J. (1990). *Essentials of Psychological Testing* (3rd ed.). New York: Harper and Row.

de Jong, R. & Westerhof J.K. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research*, 4, 51-85.

Dempo, M. & Gibson, S. (1985). Teachers' sense of efficacy: An important factor in school achievement. *The Elementary School Journal*, 86, 173-184.

Den Brok, P., Levy, J., Wubbels, T., & Rodriguez, M. (2003). Cultural influences on students' perceptions of videotaped lessons. *International Journal of Intercultural Relations*, 27, 268-288.

Dorman, J.P. (2003). Cross-National Validation of the What Is Happening In this Class (WIHIC) questionnaire using confirmatory factor analysis. *Learning Environments Research, 6*, 231-245.

Doyle, W. (1986). Classroom Organization and Management. In M. C. Wittrock (Ed.) *Handbook of Research on Teaching* (pp. 392-431). New York: MacMillan.

Ellett, C.D., & Garland, J.S. (1987). Teacher evaluation practices in our largest school districts: Are they measuring up to "state-of-the-art" systems? *Journal of Personnel Evaluation in Education, 1* (1), 69-92.

Everston, C. M., Anderson, C., Anderson, L., & Brophy, J. (1980). Relationships between classroom behaviour and student outcomes in junior high math and English classes. *American Educational Research Journal, 17*, 43-60.

Flanders, N. (1970). *Analyzing Teacher Behavior.* Reading, MA: Addison-Wesley.

Fraser, B. J. (1994). Research on classroom and school climate. In D. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 493–541). New York: Macmillan.

Fresko, B., & Nasser, F. (2001). Interpreting student ratings: Consultation, instructional modification, and attitudes towards course evaluation. *Studies in Educational Evaluation, 27*, 291–305.

Galton, M. (1987). An ORACLE Chronicle: A decade of classroom research. *Teaching and Teacher Education*, 3 (4), 299-313.

Goh, S. C., & Fraser, B. J. (1998). Teacher interpersonal behaviour, classroom environment and student outcomes in primary mathematics in Singapore. *Learning Environments Research, 1*, 199–229.

Goldstein, H. (2001). Using pupil performance data for judging schools and teachers: scope and limitations. *British Journal of Educational Research, 27* (4), 433-442.

Goldstein, H. (2003) (3rd Edition). *Multilevel statistical models.* London: Edward Arnold.

Gray, J., Goldstein, H., & Thomas, S. (2001). Predicting the future: The role of past performance in determining trends in institutional effectiveness at A level. *British Journal of Educational Research, 27* (4), 391-406.

Griffin, G.A. & Barnes, S. (1986). Using research findings to change school and classroom practice: Results of an experimental study. *American Educational Research Journal*, 23 (4), 572-586.

Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling.* New York: The Guilford Press.

Kyriakides, L. (2001). Measurement of Teaching in Cyprus: Limitations of current practice. *Proceedings of the 4th Annual Conference of the Cyprus Educational Association.* Nicosia.

Kyriakides, L. (2005). Extending the Comprehensive Model of Educational Effectiveness by an Empirical Investigation. *School Effectiveness and School Improvement, 16* (2), 103-152.

Kyriakides, L. & Campbell, R.J. (2003). Teacher Evaluation in Cyprus: Some conceptual and methodological issues arising from Teacher and School Effectiveness Research. *Journal of Personnel Evaluation in Education, 17* (1), 21-40.

Kyriakides, L., Campbell, R.J. & Gagatsis, A. (2000). The significance of the classroom effect in primary schools: An application of Creemers' comprehensive model of educational effectiveness, *School Effectiveness and School Improvement* 11 (4), 501-529.

Kyriakides, L., Demetriou, D. & Charalambous, C. (2006). Generating criteria for evaluating teachers through teacher effectiveness research. *Educational Research.*

Leary, T. (1957). *An interpersonal diagnosis of personality*. New York: Ronald Press Company.

Loup, K.S., Garland, J.S., Ellett, C.D., & Rugutt, J. K. (1997). Ten year later: Findings from a study of teacher evaluation practices in our 100 largest school districts. *Journal of Personnel Evaluation in Education, 10* (3), 203-226.

Marsh, H.W. & Roche, L.A. (1997). Making students' evaluations of teaching effectiveness effective: the Critical Issues of Validity, Bias and Utility, *American Psychologist, 52*(11), 1187-1197.

Maruyama, G. M. (1998). *Basics of Structural Equation Modeling.* Thousand Oaks, California: SAGE.

McGreal, T.L. (1983). *Successful teacher evaluation*. Alexandria, VA: Association for Supervision and Curriculum Development.

Medley, D. (1979). The effectiveness of teachers. In P. Peterson & H. Walberg (Eds.) *Research on Teaching: Concepts, Findings and Implications.* Berkeley, CA: McCutchan.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York: Macmillan Publishing Co.

Millman, J. (Ed.) (1997). *Grading Teachers, Grading Schools: Is student achievement a valid evaluation measure?* Thousand Oaks, California: Corwin Press Inc.

Moos, R. H. (1979). *Evaluating educational environments: procedures, measures, findings and policy implications*. San Francisco: Jossey-Bass.

Muthén, L. K., & Muthén, B. O. (1999). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.

Peterson, K.D. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal, 24*, 311 – 317.

Peterson, K.D. (1995). *Teacher evaluation: A comprehensive guide to new directions and practices*. Calif.: Corwin Press, Inc.

Peterson, K. D., Wahlquist, C. & Bone, K. (2000). Student surveys for teacher evaluation. *Journal of Personnel Evaluation in Education, 14* (2), 135-153.

Redfield, D. & Rousseau, E. (1981). A meta-analysis of experimental research on teacher questioning behaviour. *Review of Educational Research, 51,* 237-245.

Rigdon, E. E. (1998). Structural Equation Modeling. In G.A. Marcoulides (Ed.) *Modern methods for business research* (pp. 251-294). Mahwah, NJ: Lawrence Erlbaum Associates.

Rose, J.A., Cousins, B.J. & Gadalla, T. (1996). Internal-teacher predictors of teacher efficacy. *Teaching and Teacher Education, 12* (4), 385-400.

Rosenshine, B. (1971). *Teaching Behaviours and Student Achievement*. London: NFER.

Rosenshine, B. & Furst, N. (1973). The use of direct observation to study teaching. In R.M.W. Travers (Ed) *Second Handbook of Research on Teaching*. Chicago: Rand McNally.

Rosenshine, B. & Stevens, R. (1986). Teaching Functions. In M. C. Wittrock (Ed.) *Handbook of Research on Teaching* (pp. 376-391). New York: MacMillan.

Sanders, W. L., & Horn, S. (1994). The Tennessee value-added system (TVAAS): Mixed methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8* (3), 299-311.

Sax, G. (1997). *Principles of Educational and Psychological Measurement.* Belmont, CA: Wadswoth Publishing Company.

Scheerens, J. & Bosker, R (1997). *The Foundations of Educational Effectiveness.* Oxford: Pergamon.

Schibeci, R. A., Rideng, I. M., & Fraser, B. J. (1987). Effects of classroom environment on science attitudes: A cross-cultural replication in Indonesia. *International Journal of Science Education, 9,* 169–186.

Schoenfeld, A. (1992). Learning to think mathematically: Problem solving, metacognition and sense making in mathematics. In D. A. Grouws (Ed). *Handbook of research on mathematics learning and teaching* (pp. 334-370). New York: MacMillan.

Scriven, M. (1994). Duties of the teacher. *Journal of Personnel Evaluation in Education, 8,* 151-184.

Scriven, M. (1995). Student ratings offer useful input to teacher evaluations. *Practical Assessment, Research & Evaluation,* 4(7).

Shavelson, R. J. (1973). What is "the" basic teaching skill? *Journal of Teacher Education,* 14, 144-151.

Simon, A. & Boyer, E. (1970) (Eds.). *Mirrors of behaviours: An anthology of observation instruments continued, 1970 supplement,* Volumes A and B. Philadelphia: Research for Better Schools.

Smith, L. & Land, M. (1981). Low-inference verbal behaviors related to teacher clarity. *Journal of Classroom Interaction,* 17, 37-42.

Smith, N.L. (1991). Evaluation reflections: The context of investigations in cross-cultural evaluations. *Studies in Educational Evaluation, 17,* 3–21.

Snijders, T. & Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* London: Sage.

Soodak, L.C. & Podell, D. M. (1996). Teacher efficacy: Toward the understanding of a multifaceted construct. *Teaching and Teacher Education, 12* (4), 401-411.

Stallings, J. (1985). Effective Elementary classroom practices. In M J Kyle (Ed). *Reaching for excellence: An effective schools sourcebooks*. Washington DC: US Governing Printing Office.

Stronge, J.H, & Ostrander L.P., (1997). Client Surveys in Teacher Evaluation. In J. Stronge (Ed.), *Evaluating teaching: a guide to current thinking and best practice* (p. 129 – 161). Calif: Corwin Press.

Stronge, J.H., Helm, V.M., &Tucker, P.D. (1995). *Evaluation handbook for professional support personnel.* Kalamazoo: Western Michigan University, Center for Research on Educational Accountability and Teacher Evaluation. In Stronge (1997)

Teddlie, C., Reynolds, D., & Sammons, P. (2000). The methodology and scientific properties of school effectiveness research. In C. Teddlie and D. Reynolds (Eds.), *The international handbook of school effectiveness research (pp. 55-133).* London: Falmer Press.

Thompson, A. G. (1992). Teachers' Beliefs and Conceptions: A syntheses of the research. In D.A. Grouws (Ed.) *Handbook of Research on Mathematics Teaching and Learning*, (pp. 127-145). New York: MacMillan.

Walberg, H.J. (1979). *Educational environments and effects: evaluation, policy, and productivity*. Berkely: McCutchan.

Walberg, H.J. (1986). Syntheses of Research on Teaching. In M. C. Wittrock (Ed.) *Handbook of Research on Teaching* (pp. 214-229). New York: MacMillan.

Wang, M.C.; Haertel, G.D. & Walberg, H.J. (1990). What influences learning? A content analysis of review literature. *Journal of Educational Research, 84* (1), 30-43.

Watzlawick, P., Beavin, J.H., & Jackson, D. (1967). *The pragmatics of human communication*. New York: Norton.

Wubbels, Th., & Brekelmans, M. (1998). The teacher factor in the social climate of the classroom. In B. J. Fraser, & K. G. Tobin (Eds.), *International Handbook of Science Education* (pp.565-580). London: Kluwer Academic Publishers.

Wubbels, T., Brekelmans, M., van Tartwijk, J., & Admiraal, W. (1997). Interpersonal relationships between teachers and students in the classroom. In H.C. Waxman & H. J. Walberg (Eds.). *New directions for teaching practice and research* (pp.151-170). Berkeley, CA: McCutchan Publishing Company.

Wubbels, T., Créton, H.A, & Hooymayers, H.P. (1987). A school-based teacher induction programme. *European Journal of Teacher Education, 10,* 81-94.

Wubbels, T., & Levy, J. (1991). A comparison of interpersonal behavior of Dutch and American teachers. *International Journal of Intercultural Relations, 15*, 1-18.

Young, P.I., Delli, D.A., & Johnson, L. (1999). Student evaluation of faculty: Effects of purpose on pattern. *Journal of Personnel Evaluation in Education*, 13(2), p.179 – 190.