Evaluation of Teaching Effectiveness in Taiwan's EFL context

Bi, Shian-yun (畢先芸)
The Language Training and Testing Center
hsienyup@gmail.com

This article explores three aspects of teaching effectiveness (TE) and its evaluation: (1) The construct of TE and its importance, (2) The importance of TE evaluation and its models/approaches, (3) Frequently used TE evaluation models and approaches in Taiwan's EFL context. In addition, an actual case is provided at the end to depict what efforts can be made to improve an evaluation scheme in order to make it more constructive to language teachers in that specific EFL context.

INTRODUCTION

Instruction is greatly determined by content. Some instructional principles or methods are interdisciplinary, while some are unique to a particular subject. The requirements of an effective math teacher, for example, cannot be exactly the same as those of an effective language teacher, although they may overlap somewhat. Therefore, the effectiveness of instruction can be more appropriately evaluated in the light of teaching context – what subject is taught, and under what circumstances it is taught.

In Taiwan, an EFL environment where English is commonly studied as an L2 and plays a significant role in both education and work, the effectiveness of English teachers is frequently discussed. To improve teaching quality, various data collection approaches are used to examine English teaching effectiveness in formal educational settings, such as public schools and colleges, as well as non-formal settings, such as private language institutes. The purposes of this article are first, to explore what models/methods have been used in Taiwan to evaluate English teaching effectiveness and to identify variables that have been insufficiently examined, and second, to give suggestions for their measurement. Finally, an actual case is examined to illustrate what variables are covered in that particular EFL setting and how the existing evaluation scheme could be improved.

THE CONSTRUCT OF TEACHING EFFECTIVENESS (TE)

Teaching effectiveness is also variously referred to as teacher effectiveness, instructional effectiveness, teaching efficiency, and teacher performance. Teaching effectiveness, simply put, is the quality of the teaching force. Anderson (2004) portrays effective teachers as goal-achieving people who want to achieve goals set by themselves, educational authorities, or school administrators (p.22). Stronge (2007), synthesizing research findings, portrays effective teachers as well-trained professionals who manage the classroom well and know how to plan, teach, and monitor student progress (p. xi).

Teaching effectiveness is a multidimensional concept that means different things to different parties. Catano & Harvey (2011) reported nine teaching effectiveness competencies identified by students: communication, availability, creativity, individual consideration, social awareness, feedback, professionalism, conscientiousness, and problem-solving. Conversely, from the perspective of educational professionals, effective teachers are those who improve student learning by enhancing several factors: student motivation, persistence, creativity, knowledge applicability, and global competency (Byrd & Rasberry, 2011). In addition, effective teachers are able to help their peers improve their expertise (ibid).

Effective English language teachers require a certain set of abilities. Coniam and Falvey (1999) illustrated the core competencies needed by effective EFL teachers, listing teaching ability, language awareness (knowledge about language), and language ability (knowledge of language). To achieve the goal of effective teaching, first of all, EFL teachers must have good teaching skills. Second, they should know quite a lot about the English language systems, how the English language works, and how people learn it and use it. Third, they must have good English language proficiency. For non-native English speakers teaching EFL classes, developing all these competencies is especially challenging because of the multiple roles the teachers need to take on – those of language teacher, language analyst, and language user (Edge, 1998).

The above analysis depicts what effective teachers are like and what qualities effective EFL teachers require. These theoretical constructs lead to the following discussion of some essential issues about TE – the importance of TE, the empirical evidence that supports that importance, how TE has been evaluated in Taiwan, and the purposes TE serves.

THE IMPORTANCE OF TEACHING EFFECTIVENESS AND ITS EVALUATION

High-quality teaching is a strong contributing factor in learning outcomes. A large-scale empirical study showed how differences in achievement gains for students are associated with teacher credentials (Clotfelter, Ladd, & Vigdor, 2007). It collected end-of-course test scores in multiple subjects taken by ninth or tenth graders in North Carolina and matched them with administrative information on teacher references. That study found several significant predictors of student achievement: teacher's experience, teacher certification test scores, regular licensure, certification of teaching skills, and academic background. In addition, students exposed to teachers with weaker credentials (i.e., no teaching experience; low teacher test scores; no regular licensure; certification, but not in the specific/a related subject; no certification by the National Board; no graduate degree) tended to have lower achievement than students exposed to teachers with stronger credentials.

In Taiwan, great efforts have been made to ensure good teaching quality. Teachers in all disciplines in elementary schools and secondary schools have been encouraged by their schools or by government authorities to undergo performance evaluations on a voluntary basis. Such an approach to evaluation has also been adopted on many university/college campuses, with a legal status supported by the University Act, which gives universities/colleges the right to "carry out self evaluation on teaching, research, services, instruction, academic affairs, administration, student participation and other proceedings; regulations for the evaluation shall be formulated by the universities" (Article 5). Actually, teaching effectiveness is evaluated not only for the benefit of teachers but also for the benefit of students and school administrators. For students, particularly college students, the evaluation results may serve as course selection references. For school administrators, the results may be useful in making promotion decisions, improving curriculum structure, and collecting data for future research.

The main purpose of evaluation is to collect information for making two types of decisions, formative and summative. These types of decisions are made based on the evidence revealed by the evaluation results, and both of them greatly affect teachers' professional lives (Berk, 2005). Formative decisions are made by teachers to improve their teaching performance, while summative decisions are made by administrators to judge their personnel. These decisions in particular need to be made on a fair basis because they have crucial importance to a teacher's career. Thus, it is important to identify appropriate evaluation models or methods that best fit different educational settings and disciplines.

MODELS OF TEACHING EFFECTIVENESS EVALUATION

The present literature and current practice show that TE is measured either by a single method or by a combination of multiple methods. Methodologically speaking, if only a one-dimensional approach, such as teacher ratings or teacher self-report, is chosen to measure teaching performance, data bias may arise. For this reason, researchers have concluded that it is better to use information drawn from multiple sources because such information provides a more comprehensive picture of teacher performance (Algozzine et al., 2004; Arreola, 2004; Braskamp & Ory, 1994; Cashin, 1995; Emery, Kramer, & Tian, 2003; Felder & Brent, 2004). In a similar vein, Berk (2005) proposed the triangulation of evaluation data in hopes that the strengths and weaknesses of different data sources could compensate for each other. He listed the following potential data providers: (1) students, (2) peers, (3) administrators or department heads, (4) instructors, (5) graduates, and (6) graduates' employers. The rationale behind the first four data sources is quite clear: Students and instructors directly involved in classroom teaching can provide first hand data, while peers and school administrators who observe classes can critique the teaching from a professional perspective.

However, the information collected from graduates or their employers functions rather differently. Graduates' opinions are retrospective in nature, focusing more on the program or the quality of the teaching force as a whole than on the teaching performance of a specific instructor. The information provided by graduate employers covers graduates' knowledge and skills they present in the workplace. This information can help identify the strengths and weaknesses of programs the graduates attended previously and thus could indicate areas needing improvement. However, since the employers are not directly involved in the teaching/learning process, employers' opinions should be used with caution.

Data gathered for the TE evaluation can be either quantitative or qualitative. Frequently used tools for quantitative data collection are rating scales, checklists, or achievement tests, while tools for quantitative data collection usually include interviews, classroom observations, and reviews of teaching materials. Questionnaires and forms can be self-constructed or ready-made. Ready-made tools developed professionally would be a good choice, since their reliability and validity have already been verified. Some ready-made questionnaires are available free of charge, such as the Ohio State Teacher Efficacy Scale (Tschannen-Moran & Woolfolk Hoy, 2001). There are also commercially-available student rating questionnaires, such as the SIR II (or Student Instructional Report) developed by ETS and the IDEA System from Kansas State University.

In Taiwan's EFL context, ready-made tools are not popular for the evaluation of English teaching effectiveness. In EFL in Taiwan, most evaluation tools are either self-constructed by the schools/researchers themselves or modified versions of the evaluation form templates developed by the Committee of Educational Research of the Ministry of Education (available online at http://140.111.34.34/main/index.php).

One value of a scheme with multiple data sources is flexibility. If an evaluation scheme has multiple data sources, each source can be weighted differently, depending on how much impact each has in the context (Arreola, 2004). If, in a particular context, expert opinions are more respected than teacher self-ratings and student ratings, an evaluation scheme that uses all three as data sources may assign 40% to peer ratings and 30% respectively to teacher self-evaluation and student ratings. Variable weight assignments can also be applied to individual measurement items. For example, if a checklist item evaluating a teacher's ability to motivate students to learn has the same three data sources mentioned above, data from teacher self-evaluations and peer ratings could be given less weight than data from student ratings, since the students know best if their enthusiasm for learning increases due to the instruction they receive.

In fact, over the past thirty years, teacher evaluation has been the most popular topic of theses/dissertations on educational evaluation in Taiwan (陳素秋, 2006). In the past ten years, the number of English teaching effectiveness research/evaluation projects has been increasing. However, evaluation itself is never a simple task. The above elaboration on how TE data can be collected and measured makes it clear how diverse and complex TE evaluation can be. With this understanding, it would be interesting to examine how English teaching effectiveness has been evaluated, since this is an issue of growing importance in Taiwan.

EVALUATION METHODS/MODELS OF ENGLISH TEACHING EFFECTIVENESS IN TAIWAN

Generally speaking, methods used for collecting TE evaluation data in Taiwan are context dependent; i.e., certain data collection tools are particularly popular in a specific education setting. Based on the current state of practice and related studies of the evaluation of English teaching effectiveness, this section examines how the most common TE data collection methods/models are employed in Taiwan. Suggestions for measurement are provided later with a critique of a real case.

Self-evaluation Ratings

Self-evaluation in the form of a rating scale is the data collection tool mostly widely adopted by empirical studies of English teaching effectiveness conducted in elementary school and secondary school settings. Most of these studies explore TE with other dependent variables, variously professional knowledge (唐偉烈, 2006; 陳佳莉, 2010), professional perception (林育妃, 2006; 劉依玲, 2006), professional growth (呂淑惠, 2005; 張雲珠, 2008; 陳鴻錦, 2011), personality traits (陳鴻錦, 2011), teaching beliefs (洪湘慧, 2011), work stress (李小娟, 2008), and classroom climate (劉依玲, 2006). These studies share three things in common. First, the research purpose is to figure out the relevance of/relationships among TE and one or two other dependent variables. Second, they use self-constructed rating scales as the only data collection tool. Third, all of the data come from teachers. Such an arrangement makes sense from the researcher's perspective, since rating scales allow a large amount of quantitative data to be gathered very quickly and conveniently. In addition, it is much less effort to collect data from one single source – teachers. Finally, as Berk (2005) has indicated, the picture of their own teaching that teachers paint is provided from an insider's perspective and thus unobtainable from other data sources. However, drawing TE data exclusively from teachers requires the researchers to presume that teachers are professionals who know quite well what good teaching should be like—a questionable presumption at best. Also, since teachers naturally present themselves in the best light, the self-evaluation ratings data, if not reflected in outside evaluations, are just biased estimates.

To look at it from another perspective, self-evaluation has its irreplaceable importance in Taiwan's EFL context, especially for English teachers who are non-native English speakers. Undeniably, native English speaking teachers (NS) win trust from learners more easily than non-native English teachers (NNS). Such a tendency is particularly strong in the commercial learning context, although NNS teachers have their own advantages – they are usually grammatically analytical and familiar with the difficulties that learners might encounter during the process of language acquisition. In this situation, NNS English teachers especially need to evaluate themselves to justify their teaching beliefs and classroom practices; their collective voice should not be ignored.

Student ratings

For more than a decade, student ratings have been the most popular evaluation tool for measuring teaching effectiveness on Taiwan's college/university campuses. Although teacher self-evaluations are used by some universities as part of their teacher evaluation schemes (e.g., National Taiwan Normal University and Chang Jung Christian University), student ratings are used by the majority of colleges and universities as their only data source for evaluating teaching effectiveness. Originally, the purpose of student ratings was for formative use; i.e., the results just served as a reference for teachers to improve their teaching. Today, however, the results are often used for the summative purpose. This change means that such evaluations have an impact on a teacher's promotion, salary, annual leave, and even the right to be funded by schools to employ part-time teaching assistants.

The use of student ratings has aroused much controversy. Although student ratings can provide important information for formative (Murray, 1984; Arubayi, 1986) and summative uses (Murray, 1984; Thorne, 1980), they have spurred many complaints, too. Some of the most common problems are biased data, provided by dishonest/unserious learners; penalties directly associated with bad evaluation results; the unavailability of customized survey forms; and unclear item wording, which may cause confusion. (周祝瑛, 2009).

To make student ratings more valuable as a measuring tool, it is important to clarify what purpose such a tool serves and what actions can be taken to improve its reliability and validity. As Berk (2005) indicated, it is fine to use data from student ratings for making both formative and summative decisions, but for the latter, drawing data from multiple sources is a must. Therefore, at least, student self-evaluation has to be added to the survey in order to screen unserious student raters, and, to avoid confusion and misunderstandings, student raters must give reasons or examples to support or substantiate their opinions. Even though these extra measures are being taken, survey results need to be interpreted very carefully since, in practice, school-wide TE evaluation projects involve many situational variables that may affect the rating results yet cannot be well-controlled, such as academic disciplines/subjects, class sizes, and student levels (Emery, Kramer, & Tian, 2003).

In the EFL context, before results are analyzed and interpreted, it is necessary to answer several basic questions that contain variables unique to English teaching: (1) Does the teacher teach in English or in Chinese? (2) Is it a lecture-based course or a language skill training course? (3) Is the same class of students shared by both NS and NNS teachers? (4) Are courses of the same level taught by both NS and NNS teachers? The coexistence of NS and NNS teachers might lead student raters to draw comparisons, which could be quite unfair to NNS teachers. With so many confounding variables involved, drawing such comparisons among teachers would be difficult and hardly helpful to instruction. Comparisons, instead, should be drawn within an individual instructor, which means that a teacher's baseline performance on a specific course should be measured and used as the evaluation basis. When summative decisions are being made, it is especially crucial to collect longitudinal data.

Peer Ratings

Peer ratings play an important role in a TE evaluation scheme that has multiple data sources. As mentioned above, teacher self-evaluations can be valid only if the results are reflected by those who are actually involved in the teaching process or are responsible for the supervision of teaching quality. One good source is peer raters, since peers can give suggestions from an expert perspective, whether on in-class teaching or on materials prepared by teachers. Peer ratings can also serve as data to complement student ratings because, unlike student raters, who might have personal biases in favor of or against their teachers, peer raters are detached from the teacher-student relationship and thus can give judgments free of personal bias. However, the results of peer ratings could also be affected by other variables, such as the academic backgrounds of peer raters and their teaching beliefs, the lengths of in-class observations, the number of visits, and the availability of qualified observers. These weaknesses of peer ratings have led to doubts about whether peer ratings are a valid method for the summative purpose (Center for Teaching and Learning of the University of Minnesota, 2011). In a model proposed by Brent and Felder (2004) covering multiple types of data – student ratings, peer ratings, self-ratings, and administrator ratings – peer ratings can be used for summative purposes if two raters work together. However, even with two peer raters, it is still questionable for such a small group of people to make decisions with far-reaching impacts on a teacher's career, a very important aspect of life.

In Taiwan's EFL setting, getting feedback from peers is strongly encouraged by the Ministry of Education (also known as the MOE) for use as a TE evaluation method at elementary and secondary schools. Usually, such evaluations are done by the teachers' colleagues at the same school, though sometimes they are done by university professors on behalf of the MOE. Interestingly, the MOE has stated clearly that the observation results are used for the formative purpose only, since so many teachers worry about their job security.

In the interpretation of the peer rating data, a key variable should be the years of teaching experience of the teacher being rated. A good example is the in-class observation template provided by the MOE. This form includes many core aspects of teaching, such as curriculum design, instruction skills, and classroom management. Nevertheless, there is one more factor to keep in mind: The demands for a novice teacher differ from the requirements of an experienced teacher. This difference is true in all disciplines, including language learning. The teaching quality of an experienced language teacher should differ from that of a novice. Therefore, different evaluation criteria should be established for language teachers with different levels of experience. Results need to be carefully interpreted, especially when one observation form serves all teachers.

Parent Ratings

Unlike the above-mentioned measures, parent ratings are rarely discussed in the literature. However, in Taiwan's EFL setting, parent satisfaction has great importance, especially in private schools. One such example is private elementary schools in Taiwan. These schools generally allocate more hours to English teaching than do public schools, which is an important incentive for parents to send their children there. As a result, private schools collect feedback from parents to gauge the parents' impressions of their English teaching program and to find ways to further satisfy the parents.

However, from the methodological perspective, such data have some obvious flaws. First, it is second-hand data; i.e., parents are not actually involved in the teaching or learning process. Second, parents are very possibly unqualified raters, since most parents have limited knowledge of what good English teaching should be like. The third flaw is that their opinions might be affected by their children's attitudes. Such weaknesses are very dangerous for a TE evaluation scheme. If schools do want parents to play a role in the TE evaluation scheme, parents are able at best to provide some facts about their children's English learning at home, such as signs of interest in learning English that their children show when they are with their parents. In short, information gathered from parents should have, at most, minor importance in any TE evaluation scheme.

Given the strengths and weaknesses of the TE evaluation models and methods above, an actual case is provided here. This case is used to illustrate current practices, describe limitations of that particular context, and propose possible methods to evaluate teachers on a more equal basis.

AN ACTUAL CASE CRITIQUE

Case Description

A private language training institute, one which provides language courses throughout the year, evaluates all teachers with student ratings. Teacher evaluations take place for primarily a formative purpose once around the middle of the term, usually in week 5 or week 6. Except for writing classes and specially-designed classes, for which surveys with extra subject-specific items are used, all language courses are surveyed with a uniform survey. Classroom observations, on the other hand, are done mostly for novice or newly-hired teachers, and all are done by program coordinators. Due to the heavy workload of the coordinators, other teachers are observed only when they are teaching new books, when the student rating surveys indicate problems, or when the students file complaints. Low survey scores often lead to a group interview with the learners, and when a consistently high-performing teacher receives low survey scores, the results are considered with caution. Teachers are given training when specific problems are identified. The teaching staff is composed of both NS and NNS of the target languages. Different combinations of teachers may share a half-day English class (3 learning hours per day with 2 or 3 subjects) or an intensive English class (6 learning hours per day with 5 subjects). Normally, class size ranges from 8 to 16 students, and the vast majority of students are adults.

Analysis and Improvement of the Evaluation

The timing of the survey and the formative focus allow training to be provided when problems are detected. In this sense, the TE practices do help teachers to improve their performance. However, in general, this language institute has two evaluation tracks: (1) multiple measures (classroom observation and student ratings) for novice/new teachers, and (2) a single measure (student ratings) for more experienced teachers. Such a single measure design may expose experienced teachers to a higher risk of inaccurate evaluation. Since the manpower shortage does not allow classroom observations of all of the teachers, some remedial measures need to be taken to make up for the inequity.

One convenient solution might be to have teachers do self-evaluations and have course coordinators review their teaching plans and materials. Another possibility is to employ baseline measurement, a highly flexible and informative method. Though it is currently employed in the language institute above, it could be more widely and systematically applied either to individuals or to groups of teachers. An individual teacher's baseline shows his or her teaching quality over time, or longitudinal performance; thus, any result that is significantly different from the baseline longitudinal performance can be easily detected. Baseline measurement can also be applied to teachers of a particular course, teachers of a particular level of a course, teachers with different cultural backgrounds (NS or NNS), or teachers of different experience levels. However, it must be noted that such baseline information from groups should not be used for comparisons. Instead, it should serve only as the baseline for an alarm that alerts the program coordinators/supervisors to problems so that they can intervene before the situation worsens.

Other survey techniques can also be applied to improve the existing evaluation scheme:

- Set different standards for teachers of different experience levels;
- Avoid surveying a teacher immediately after teaching;
- Survey teachers on separate days to avoid the contrast effect; and/or
- Discard quantitative data collected from a very small sample due to the low statistical power, and instead collect qualitative data.

CONCLUSION

Establishing a teacher evaluation system that brings the greatest benefits to learners, teachers, and school administrators is a challenging mission. Some context-specific limitations cannot be overcome easily, and unexpected flaws in a TE evaluation scheme always exist, no matter how well the scheme is designed. Nevertheless, any TE evaluation scheme should be both outcome- and improvement-oriented so that students achieve better learning and teachers advance their expertise. Such a scheme will eventually lead to the formation of an everlasting virtuous circle.

REFERENCES

張銘峰（2005）。＜國中英語教師專業成長與教學效能關係之研究＞。臺中師範學院國民教育研究所，未出版，臺中市。

劉依玲（2006）。＜教育改革下國小英語教師專業認知與班級氣氛、教學效能之關係研究＞。高雄師範大學教育學系，未出版，高雄市。

呂淑惠（2005）。＜國民小學英語教師專業成長與教學效能之研究－以屏東縣為例＞。屏東師範學院行政教育研究所，未出版，屏東市。

李小娟（2008）。＜桃園縣國中英語教師工作壓力與教學效能之關係研究＞。中原大學教育研究所，未出版，桃園縣。

張雲珠（2008）。＜國民中學英語科教師專業成長與教學效能之研究-以台中縣為例＞。國立彰化師範大學教育研究所，未出版，彰化市。

陳素秋（2006）。＜台灣近三十年來教育評鑑研究生論文分析＞。《教育政策論壇》，第九卷第三期。

陳佳莉（2010）。＜高職英文科教師知覺專業知能與教學效能之相關研究＞。國立彰化師範大學工業教育與技術學系，未出版，彰化市。

林育妃（2006）。＜教育改革下國中英語教師專業認知與教學效能之關係研究—以南部地區為例＞。國立高雄師範大學教育學系，未出版，高雄市。

教育部教育研究委員會 (2008)。《試辦中小學教師專業發展評鑑宣導手冊》。http://140.111.34.34/docdb/files/dma7db0401131e2d0e9.pdf

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., Mohanty, G.., & Spooner, F. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching, 52*(4), 134-141.

Anderson, L. W. (2004). Increasing teacher effectiveness. *Fundamentals of Educational Planning, 79*. Paris: UNESCO.

Arreola, R. A. (2004). Developing a comprehensive faculty evaluation system. http://www.magnapubs.com/images/041109AC.pdf.

Arubayi, E. (1986). Students' evaluations of instruction in higher education: a review. *Assessment and Evaluation in Higher Education, 11*, 1-10.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and learning in High Education, 17*(1), 48-62.

Braskamp, L.A., & Ory, J. C. (1994). *Assessing faculty work.* San Francisco: Jossey-Bass.

Byrd, A., & Rasberry, M. (2011). *Teacher and teaching effectiveness: A bold view from National Board Certified Teachers in North Carolina*. North Carolina: CTQ center for teaching quality.

Catano, V. M., & Harvey, S. (2011). Student Perception of Teaching Effectiveness: Development and Validation of the Evaluation of Teaching Competencies Scale (ETCS). *Assessment & Evaluation in Higher Education, 36*(6), 701-717.

Cashin, W. E. (1995). Student ratings of teaching: the research revisited. *IDEA Paper No. 20, ED 302 567.*

Center for Teaching and Learning of the University of Minnesota. (2011). Peer observation guidelines and recommendations. http://www1.umn.edu/ohr/teachlearn/resources/peer/guidelines/.

Clotfelter, C., Ladd, H., & Vigdor, J. (2007). *Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects*. http://www.caldercenter.org/PDF/1001104_Teacher_Credentials_HighSchool.pdf

Coniam, D., & Falvey, P. (1999). Setting standards for teachers of English in Hong Kong–The teachers' perspective. *Curriculum Forum, 8*(2), 1-27.

Edge, J. (1988). Applying linguistics in English language teacher training for speakers of other languages. *ELT Journal, 42*(1), 9-13.

Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: a critique of student evaluations of teaching effectiveness. *Quality Assurance in Education 11*(1), 37-46.

Felder, R. M., & Brent, R. (2004). How to evaluate teaching. *Chemical Engineering Education, 38*(3), 200-202.

Brant, R. & Felder, R. M. (2004). A protocol for peer review of teaching. Session 3530. *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition.* Salt Lake City, Utah.

Gage, N. L. (1984). What do we know about teaching effectiveness? *The Phi Delta Kappan, 66*(2), 87-93.

Goddard, R. D., Hoy, W. K., & Hoy, A. W. (2000). Collective teacher efficacy: Its meaning, measure, and impact on student achievement. *American Educational Research Journal, 37*(2), 479-507.

Murray, H. G. (1984). The impact of formative and summative evaluation of teaching in North American University. *Assessment and Evaluation in Higher Education 9*(2), 117-132.

Stronge, J. H. (2007). *Qualities of Effective Teachers,* (2nd ed.). VA: Association for Supervision and Curriculum Development.

Thorne, G.D. (1980). Students ratings of instructors: from scores to administrative decisions. *Journal of Higher Education, 51*, 207-214.

Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing and elusive construct. *Teaching and Teacher Education, 17*, 783-805.

Wright, T. (2002). Doing language awareness: Issues for language study in language teacher education. In H. Trappes-Lomax & G. Ferguson (Eds.), *Language in language teacher education* (pp. 113-130). Amsterdam: John Benjamins.