

Running head: Teacher Assessment Instrument

Developing a teacher assessment instrument:

Validation and implications

Linda S. Gilbert, Ph.D.

315 Aderhold Hall

University of Georgia

Athens, GA 30602

gilbertl@uga.edu

Stephen E. Cramer, Ph.D.

Associate Director

Georgia Center for Assessment /

Test Scoring and Reporting Services

University of Georgia

1985 New Jimmy Daniel Road

Athens, GA 30602-9593

Developing a teacher assessment instrument: Validation and implications

Introduction

The standards movement for teacher preparation and professional development is scarcely new, though now driven to new levels by No Child Left Behind (NCLB) legislation (Campbell, Kyrakides, Muijs, & Robinson, 2004; Cochran-Smith, 2003; Harris, 2002; Kupermintz, 2003; Muijs, 2004; Porter, Youngs, & Odden, 2001; Lee S. Shulman, 1988a, 1988b, 1992; Stronge, 1997). The better descriptions of teaching recognize that while content knowledge is a necessary component of good teaching, it is far from sufficient. These standards go beyond simplistic content-oriented metrics. However established, standards are frequently translated into assessment or evaluation tools.

The purpose of this paper is to describe a study designed to evaluate a self-assessment instrument for teaching. This instrument is based on a framework for accomplished teaching developed as part of a larger grant for improving teaching quality. The underlying framework has been adopted throughout the state the Committee on Quality Teaching (CQT), a group of educational agencies at multiple levels, including the state teacher certification agency, the Board of Regents, and the Department of Education. The tested version is a self-assessment for teachers using a rubric-based scale and the following categories of expertise: Content and Curriculum, Knowledge of Students and their Learning, Learning Environments, Assessment, Planning and

The research reported in this paper has been developed in conjunction with the Georgia Systemic Teacher Education Program (GSTEP). GSTEP is funded by the United States Department of Education (Grant Number P336B000009), by the State of Georgia, and by the University System of Georgia. The opinions expressed in this paper are those of the author(s) and do not necessarily reflect the views of GSTEP or its funding agencies.

Instruction, and Professionalism. The primary goal of the study was to assess the validity of the instrument, particularly as it relates to showing differences in teacher development.

The goal of the analysis was to answer several technical questions about the data produced by the instrument. These questions included:

1. Does the instrument generate variance across candidates (i.e., do different groups look different)?
2. Do individual categories have appropriate item characteristics, such as frequency distribution of responses and Indicator-Category correlation?
3. Do the categories represent empirical as well as conceptual dimensions, when explored through a factor analytic approach?

This paper provides affirmative answers to these questions, although the category analyses (question 3) showed some qualitative differences based on respondent type.

Perspectives or theoretical framework

The developers of the framework underlying this assessment instrument recognize that “good teaching” is both contextual and variable from teacher to teacher (Campbell et al., 2004; Lee S. Shulman, 2004). In creating both the original framework and the self-assessment tool based on it, they primarily sought to support teacher preparation, professional development, and reflection on practice, recognizing both the pitfalls inherent in teacher evaluation and assessment and the political forces necessitating such activities (Porter et al., 2001; Serpell, 2000; Lee S. Shulman, 1988a, 1992; Smith & Ingersoll, 2003; Wang & Odell, 2002). Given the diversity of teacher preparation and abilities, this instrument is intended to provide a means of developing personalized professional development goals based on individual needs.

Context and background

The instrument used in the study was based on an underlying framework. This framework (originally called the GSTEP Framework for Accomplished Teaching, now the Georgia Framework for Teaching) was initially developed as part of the Georgia Systemic Teacher Education Program (Georgia Systemic Teacher Education Program (GSTEP)). GSTEP work began in 2000, supported by grant from the U.S. Department of Education, with further funding from the State of Georgia and The University System of Georgia Board of Regents. For the sake of clarity, the GSTEP/Georgia Framework for Accomplished Teaching will simply be referred to as “the *Framework*” throughout this paper, though name transitions will be noted as relevant.

The GSTEP grant’s goals were to (1) establish seamless, high quality learning opportunities and support for beginning teachers (BTs), especially in Georgia’s high-need schools; (2) prepare BTs to bring all learners to high levels of achievement; and (3) create systems for policy, professional development, dissemination, and evaluation. Under Goal 1, “beginning teacher” was conceptualized as extending from the initial decision to become a teacher through at least the second year of teaching. This conceptualization was in contrast to the fragmented condition of teacher preparation at the time the grant began.

To work toward the goal of “seamless” learning and support for beginning teachers, GSTEP formed collaborative teams to revise teacher preparation curriculum, examine current induction practice, and propose initiatives addressing the larger vision. These teams included an array of stakeholders: representatives from the participating universities (both Arts and Sciences and College of Education faculty) as well as P-12 faculty and administrators from eleven collaborating school districts. Representatives

from state agencies involved with teacher preparation were also involved with GSTEP work.

In 2001, the teams specifically focusing on induction were charged with exploring ways to promote integration and support for new teachers as they transitioned from the classroom to school settings. Their combined efforts led to the first drafts of the Framework. Hudson-Ross (2005) chronicles the full development process. The work began in the spring of 2001 with a review of induction literature and a series of focus groups. Specifically, the teams reviewed a number of state models (Connecticut, Washington, California, and others) as well as induction projects such as Danielson's (1996) *Pathwise* program (Danielson, 1996) and standards for teachers proposed by National Board for Professional Teaching Standards (NBPTS) and the Interstate New Teacher Assessment and Support Consortium (INTASC). In the same time period, focus groups were held with numerous educators to explore how they would define "good teaching." At the end of this broad data collection effort, the 20 cross-institutional team members met for a day to share and analyze the insights gleaned from the focus groups and literature.

The original hope was to find an existing framework compatible with Georgia's educators' beliefs about "good teaching" and suitable for sustained implementation. Analysis identified some cross-cutting ideas and themes, some of which later developed into the GSTEP Principles (Georgia Systemic Teacher Education Program (GSTEP)). However, though the group agreed that a framework would help provide a focal point for integration of teaching preparation, the existing frameworks either seemed to lack

elements deemed critical or to be too costly to sustain (e.g., Danielson's Pathwise program).

In May 2001, a cross-institutional writing team met for three days to develop a plan “based on the literature, focus group insights, and their own experience and expertise” (Hudson-Ross, 2005). They first developed the GSTEP Principles (Georgia Systemic Teacher Education Program (GSTEP)) to guide the development of the Framework, Principles which provided a vision for much of GSTEP work from that point on. Though elaborated in the original document, the Principles can be briefly summarized as follows:

- **The Process Principle:** Learning to teach is a career-long process of development and growth.
- **The Support Principle:** All educators share responsibility for supporting their colleagues as professional peers.
- **The Ownership Principle:** Teachers design their own career paths.
- **The Impact Principle:** Effective teaching yields evidence of student learning and achievement.
- **The Equity Principle:** All students and teachers deserve equally high expectations and support.
- **The Dispositions Principle:** Productive dispositions positively affect student learning, teacher growth, and school climate.
- **The Technology Principle:** Technology facilitates teaching, learning, community building, and resource acquisition.

Initially, these Principles were intended to guide the development process, serving as “touchstones” as the Framework was written. However, they were considered so important by the developers that after the initial drafts of the Framework were completed, the Principles were frequently published in conjunction with it.

At the end of the team’s work, they had identified six domains of practice (known as “elements”) with three more subsumed into those six. The final six elements were: (1) Content and Curriculum, (2) Knowledge of Students and their Learning, (3) Learning Environments, (4) Planning and Instruction, (5) Assessment, and (6) Professionalism. Technology, diversity, and student learning were expected to pervade all six. These elements formed the core of the Framework from its inception.

After several drafts and rounds of editing, the initial *GSTEP Resource Framework* (a 12-page document) and the *GSTEP Principles* were complete. These documents were reviewed in a second round of focus groups designed to gather input across the state. Over 60 focus groups were conducted by moderator/note-taker teams in Fall 2001. Data were analyzed in a day-long retreat. Hudson-Ross (2005) summarizes the conclusions and the action taken: “Across all groups, they found a high degree of excitement about and support for the idea of a framework to guide induction, along with a good deal of concern: first that such a framework could become a mode of summative evaluation rather than remaining a teacher-directed professional development tool (as per the ownership principle), and secondly, that the framework was so highly detailed as to seem overwhelming, even to highly accomplished teachers. Based particularly on the latter finding, and relying on a range of specific comments about strengths and weaknesses in the framework, on April 3-4, 2002, a revision team reduced the 12- page document to two

pages and six elements... The document called the *GSTEP Framework for Accomplished Teaching* was then ready to be used.” (Hudson-Ross, 2005).

The final document retained the six “elements” previously listed (content and curriculum; knowledge of students and their learning; learning environments; planning and instruction; assessment; and professionalism), along with a total of 41 “indicators” across those six elements. (See attached one-page summary.) After completion, GSTEP leaders also aligned the Framework to existing state and national standards for teaching (Hertzog, 2006) The goal was not to supplant more specific standards, but to provide a “common language” across the different stakeholder groups involved in teacher preparation and practice: P-12 schools, Arts and Sciences, and Colleges of Education. (For one example, see Georgia Systemic Teacher Education Program (GSTEP), 2006)

Instrument development

Throughout the development of the *Framework*, concerns and tensions about its intent and use surfaced. (Sztajn et al., 2006) The goal of the Framework was to support teacher development, not teacher evaluation. However, it was almost inevitable that assessment instruments would be desired by teacher educators and others responsible for teacher preparation and development.

Thus, the Framework served as the basis for the development of the “Self-Assessment for Accomplished Teaching,” usually known as the “Self-Assessment Tool” This instrument was developed by a committee of faculty centered at Valdosta State University, one of the partner institutions. They developed and pilot-tested several variants of the instrument and associated scoring rubrics and scales. During pilot-testing, they administered versions of the Self-Assessment Tool online to their own institution’s

students, both undergraduate and graduate, and held focus groups with student teachers, university faculty and induction leaders for additional feedback. (Hertzog, Monetti, Minor, & Judd, 2004, February). The version that emerged from this development process served as the basis for the current study.

In the version tested, descriptive rubrics for each indicator accompanied a numerical rating scale that ranged from 1 (lowest) to 5 (highest). The option of “not able to rate” (NATR) was also provided. For each of the indicators, the descriptors corresponding to the 1-5 scale were unique, i.e., the instrument was in the form of a true rubric and not a simple rating scale.

Methods and data sources

This study was conducted by an evaluation and testing department affiliated with the lead university but not connected with the funding grant, in collaboration with the internal evaluator for GSTEP (the lead author). In order to obtain a sample of respondents, the instrument was developed into an online survey using 41 questions (the Framework “indicators”) divided into the six constructs or categories (the Framework “elements”), plus a section on demographic information and a space for open-ended comments on the instrument itself. Teachers and teacher educators in the partner institutions and their affiliated school systems were invited via email to participate in the survey, and to recommend or require that their students also complete the instrument, either for their own professional development or as a requirement for class or practice teaching. Participants could opt out of the study at any point, including after completion, following the strict IRB guidelines for online surveys. Participants did not receive any compensation from GSTEP for participating. We encouraged them to print out their

response summary for use in personal planning and in consultations with a supervisor, supervising teacher, or faculty member, but data were not collected as to whether or how often this happened.

The following analyses are based on 353 completed self-assessments collected between March 31, 2005 and December 1, 2005. However, not all of the demographic information was completed by all respondents, so frequencies noted do not add to 353.

Of the 353 completed self-assessments collected, the sample was evenly split between practicing and pre-service teachers, with a small (N=12) group teachers who self-identified as National Board Certified teachers. The characteristics of the sample provided a reasonably good fit with the general in-service and pre-service teacher population.

Results

The data were used to answer several questions about the instrument. The aspects of interest were whether we could perceive differences between teachers at different levels of preparation, and whether the constructs grouped appropriately. These developed into the three questions previously identified, and discussed below.

Question 1: Does the instrument generate variance across candidates (i.e., do different people look different)?

To answer this question, we calculated the mean, standard deviation, and frequency distribution of each of the 41 items (“Indicators”). The modal rating was 4 for all items, and the median rating was 4 or 5, indicating a high level of perceived accomplishment. We ran a set of frequency distributions of score on each Indicator by Respondent Type (preservice and practicing). The results showed clear response

differences between pre-service and practicing teachers that were significant at an alpha of .05 for all Indicators except “III-A. Learning Community.” (See Table 1.) Pre-service teachers were more likely to rate themselves lower than practicing teachers, but their median response was still a 4 on the 1 to 5 scale, which seems somewhat inflated. In other words, at least half of the sample of both pre-service and practicing teachers considered themselves to be quite well accomplished.

Table 1: *Comparison of practicing and pre-service teachers by category, all comparisons significant at alpha=.05*

Category	Practicing	Pre-service
Content and curriculum	4.19	3.79
Knowledge of students and their learning	4.06	3.69
Learning environments	4.02	3.66
Assessment	3.95	3.61
Planning and Instruction	4.10	3.77
Professionalism	4.28	3.81

Not all Indicators received 353 ratings, since one of the response options was “Not Able to Rate.” Mean ratings ranged from a high of 4.37 on Indicator VI-C, “Codes of professional conduct”, to a low of 3.42 on Indicator IV-B, “Use of pre-assessment data”. Standard deviations ranged from .72 to 1.08. All Indicators except for 2, 6, 7, 25, 32, 37, 38, and 39 had values from 1 to 5; the smallest value for these items was 2. The distribution of responses was uniform across Indicators, being generally negatively skewed, i.e., very small numbers of ratings of 1 and 2.

Seven Indicators showed NATR rates of greater than 10%. In all cases, the great majority of NATR ratings came from pre-service teachers, suggesting that these items

assess an aspect of teaching that they have not yet experienced, and not a flaw in the items themselves. Examination of the items reinforced this hypothesis.

In short, there was variation across participants: pre-service teachers do rate themselves lower than experienced teachers. Thus, the instrument appears to generate acceptable levels of variance on all items. However, the distribution of responses is quite skewed, and the self-assessments seem quite high for both practicing and pre-service teachers. These findings suggest that the instrument does have content validity, but suggests additional attention to the scale/rubric and the anchors provided to various ratings. That is, more concrete anchors might assist teachers in self-assessing more accurately, resulting in greater variance overall

We also correlated categories with Years of Experience. (See Table 2.) The initial correlations were low, due to the inclusion of the pre-service teachers, which skewed the distribution severely. We removed all pre-service teachers and all respondents with less than one year's experience. The resulting correlations, while significant, are still not strong. Years of Experience correlated best with Content and Curriculum indicators.

Table 2: *Correlations of years of experience and category scores*
All correlations significant at alpha = .05.

Category	Years of Experience
Content and Curriculum	0.45
Knowledge of Students and their Learning	0.25
Learning Environments	0.30
Assessment	0.29
Planning and Instruction	0.39
Professionalism	0.31

Question 2: Do individual Indicators have appropriate item characteristics, such as Indicator-Category correlation?

Each of the Indicators is nested within a category, or element. A testable hypothesis is that the items within a category should correlate with a subscale constructed as the sum of those responses. Accordingly, we calculated a score for each construct and correlated this score with each indicator. In all cases, the item correlated significantly with its category. We also correlated the categories with each other, and found correlations ranging from .44 to .69, indicating that while the categories are related, as one would expect, they appear to be assessing somewhat different aspects of accomplished teaching.

We further calculated Cronbach's alpha statistic for each category. (See Table 3.) Alpha coefficients ranged from 0.80 to 0.89, indicating that the category scores have good reliability in the context of self-assessment feedback.

Table 3: *Cronbach alpha reliability by category*

Category	Alpha
Content and curriculum	.80
Knowledge of students and their learning	.85
Learning environments	.85
Assessment	.89
Planning and Instruction	.87
Professionalism	.87

The category scores are further validated by comparing the responses of pre-service and practicing teachers. A test of the hypothesis that practicing teachers do

perceive themselves, as a group, as more accomplished than pre-service teachers shows that in all cases, practicing teachers have significantly ($p < .05$) higher category scores than pre-service teachers.

Question 3: Do the categories represent empirical as well as conceptual dimensions, when explored through a factor analytic approach?

To explore this question, responses from all respondents to the 41 Indicators were factor analyzed using principal axis factoring. The solution yielded six factors, based on the eigenvalue-greater-than-one principle, and was rotated using the Oblimin method with Kaiser Normalization to maximize the differences among factors. These six factors account for 60.1% of the total variance. The factor solution yielded 5 very clear factors, which might be labeled Planning & assessment, Knowledge and learning, Professionalism, Concerns beyond the classroom, and Knowledge and skills. One additional factor did not seem to lend itself to easy description.

The fact that the empirical factor structure and the *a priori* categories do not match completely does not necessarily indicate a need to reorganize the assessment. However, this is an indication that developers should look carefully at the organization of the instrument to ensure that the categories possess a high level of conceptual integration.

In an effort to explore the underlying structure of the assessment more fully, we performed separate factor analyses for more experienced and less experienced (i.e., pre-service) teachers. The identifiable factors derived from this analysis showed some significant differences. Less experienced teachers' responses grouped into only 3 identifiable areas: Understandings, Professionalism, and Classroom activities. More experienced teachers' responses grouped into 5 identifiable areas: Assessment and

learning environment, Professionalism, Planning, Content, Students and learning, and Managing and motivating. This suggests that the instrument is capturing an expected outcome: that experienced teachers have more complex and better-defined conception of teaching than those who are just beginning in the profession.

Additional comments

Additional participant comments indicate that participants generally found the instrument easy to use, though there were some concerns about wording and length. In addition, participants were concerned about potential use beyond self-assessment.

Summary

In summary, in terms of the instrument content, the items composing each of the six categories had good correlations (Cronbach's alpha coefficients ranged from 0.80 to 0.89). However, the three factor analyses of items (one for all teachers, one for pre-service, one for experienced) showed varying patterns. The factor analyses suggest that experienced teachers have a more complex view of teaching than beginners. A logical next step would be to revisit the organization of the constructs using the factor analysis to inform the organization further.

In terms of the instrument scale, though the responses were significantly different across participant types, participants rated themselves very highly overall. This suggests attention to the scale/rubric and the anchors provided to various ratings. (Note: This work is in progress.)

In terms of the validity of the instrument, the four traditional kinds of validity (construct, content, face, and criterion-related) need to be discussed separately. Construct validity, which is hypothesis-driven, was the primary focus of this study. The study did

confirm construct validity as described by the research questions. Content validity addresses the accuracy of the content, which was previously established by the expertise brought to the instrument construction. Face validity involves the reactions of the participants, which was also incorporated into the development process and confirmed by participant comments from this study. Criterion-related validity addresses the comparative or predictive power of the instrument (that is, does it correlate to other measures OR does it predict later performance). There is currently no comparative data with which to assess criterion-related validity. Criterion-related validity is one possible avenue for future study.

Educational or scientific importance of study

This study lays an important foundation for using this instrument, variations developed from it, or other self-assessments and frameworks similar to it to help teachers self-assess their teaching skills.

In terms of planning future work with teachers and research based on this pilot, the study data supports additional descriptive work currently underway on the original Framework. For example, the noticeable skew in reporting is being addressed with more concrete descriptors and the development of layered rubrics for teachers at various levels of development and expectations. These layered rubrics are envisioned in an online branching format to reduce the length of the instrument for the teacher completing it.

In terms of development, a careful examination of the factor analysis categories may be in order. The different perspectives that pre-service and experienced teachers bring to teaching are of interest in themselves. While it may be argued that the Framework is conceptualized as applying across the length of a teaching career, an

awareness of the different ways that teachers conceptualize their work would inform its use.

In terms of future research, a particularly interesting avenue to be explored further is the actual use of the instrument as a self-assessment and the consequences of use. Asking teachers using the survey how they have applied or could apply these standards to their practice would be one way to address this issue. One future possibility lies in examining the actual use of the instrument through Messick's Unified Validity Framework (Messick, 1989). Messick proposed additional validity criteria that examined relevance and utility, value implications, social consequences, and interpretation. However, he viewed construct validity as an essential first step to examining these aspects of an instrument-in-use. This study lays the foundation for the use of this instrument, and invites additional reflection on its purpose.

Similar studies with other instruments should also be conducted in order to improve our understanding of teachers, teaching and self-assessment of teaching at various career stages. An improved definition of "quality teaching" would benefit teacher preparation programs, teachers at all levels, and ultimately the students they teach.

References

Campbell, J., Kyriakides, L., Muijs, D., & Robinson, W. (2004). *Assessing teacher effectiveness: Developing a differentiated model*. New York: Taylor & Francis.

- Cochran-Smith, M. (2003). Assessing assessment in teacher education. *Journal of Teacher Education*, 54(3), 187-192.
- Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, Va.: Association for Supervision and Curriculum Development.
- Georgia Systemic Teacher Education Program (GSTEP). *Guiding Principles of the Georgia Framework for Teaching*. Retrieved March, 2007, from http://www.teachersbridge.org/library/GFT_Principles.pdf
- Georgia Systemic Teacher Education Program (GSTEP). (2006). Overview of the Integrated Standards: Alignment of the Georgia Framework for Teaching to Selected State and National Standards, *Available from The BRIDGE*, <http://www.teachersbridge.org>.
- Harris, L. (2002, February 23-26, 2002). *Standards Based Program Assessment: Connecting Teacher Performance to Student Learning*. Paper presented at the American Association of Colleges for Teacher Education, New York, NY.
- Hertzog, P. (2006). *Georgia Integrated Standards Document*. Retrieved March, 2007, from <http://www.teachersbridge.org/library/5%2BGSP%2Bstandards%2Balign.pdf>.
- Hertzog, P., Monetti, D. M., Minor, L. C., & Judd, D. (2004, February, 2004, February). *Using self-assessment to improve teacher performance*. Paper presented at the 56th annual meeting of the American Association of Colleges for Teacher Education, Chicago, Illinois.

- Hudson-Ross, S. (2005). History of the Development of the GSTEP Principles and Framework for Accomplished Teaching (Now the Georgia Framework for Teaching), *Teacher Quality Enhancement (TQE)*. Washington, DC: Available from The BRIDGE, <http://www.teachersbridge.org>.
- Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis*, 25(3), 287-298.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan
- Muijs, D. (2004). Measuring teacher effectiveness. In D. Hopkins & D. Reynolds (Eds.), *The Learning Level*. London: RoutledgeFarmer.
- Porter, A. C., Youngs, P., & Odden, A. (2001). Advances in teacher assessments and their uses. In V. Richardson (Ed.), *Handbook of Research on Teaching* (4th ed., pp. 259-297). Washington, D.C.: American Educational Research Association.
- Serpell, Z. (2000). *Beginning teacher induction: A review of the literature* (No. ED443783). Washington, DC: American Association of Colleges for Teacher Education.
- Shulman, L. S. (1988a). The paradox of teacher assessment. In *New Directions for Teacher Assessment: Proceedings of the 1988 ETS Invitational Conference*. Princeton, N.J.: Educational Testing Service.
- Shulman, L. S. (1988b). A union of insufficiencies: Strategies for teacher assessment in a period of educational reform. *Educational Leadership*(November), 36-39.

- Shulman, L. S. (1992). Research on teaching: A historical and personal perspective. In F. K. Oser, A. Dick & J.-L. Patry (Eds.), *Effective and Responsible Teaching* (pp. 14-29). San Francisco: Jossey-Bass.
- Shulman, L. S. (2004). *Wisdom of practice: Essays on teaching, learning, and learning to teach*. San Francisco: Jossey-Bass.
- Smith, T. M., & Ingersoll, R. M. (2003). *Reducing teacher turnover: What are the components of effective induction?* Unpublished manuscript.
- Stronge, J. H. (Ed.). (1997). *Evaluating teaching: A guide to current thinking and best practice*. Thousand Oaks, CA: Sage.
- Sztajn, P., Moore, J., Gilbert, L., Stitt-Gohdes, W., Hudson-Ross, S., & Hertzog, P. (2006). Development, uses, and impact of a Framework for Accomplished Teaching, Available from The BRIDGE, <http://www.teachersbridge.org>.
- Wang, J., & Odell, S. J. (2002). Mentored learning to teach according to standards-based reform: A critical review. *Review of Educational Research*, 72(3), 481-546.