

Running head: GENDER STEREOTYPE

**An Empirical Examination of Gender Stereotype from
the Result of National Board Certification**

Tracy W. Smith
Department of Curriculum and Instruction
202E Edwin Duncan Hall
Appalachian State University
phone: (828)262-2274
fax: (828)262-2686
E-Mail: smithtw@appstate.edu

Jianjun Wang
Department of Advanced Educational Studies
California State University, Bakersfield
9001 Stockdale Highway
Bakersfield, CA 93311-1099
Phone: (661) 664-3048
Fax: (661) 664-2479
E-Mail: jwang@csub.edu

J. Steve Oliver
Department of Science Education
Aderhold Hall
University of Georgia
Athens, GA 30602
Office Phone: (706) 542-4648
Fax: (706) 542-1212
E-Mail: soliver@coe.uga.edu

Paper presented at the 2005 annual meeting of the American Educational Research Association (AERA) in Montreal, Canada.

**An Empirical Examination of Gender Stereotype from
the Result of National Board Certification**

Abstract

The National Board of Professional Teaching Standards (NBPTS) is designed to recognize accomplished teachers in the profession. Validity of the national board certification hinges on a fundamental concern whether confounding factors other than teaching performance have contributed to the certification outcome. In particular, gender stereotypic influence is examined in this study using a large-scale national database in four subject areas. Besides confirming gender inequity in the scoring outcomes, the results also suggest that the outcome difference was subject-specific. Male applicants outperformed their female counterpart in science, despite the stereotypic view of teaching as a female occupation. On the other hand, female applicants consistently received higher scores in so-called non-masculine subjects, such as English and social studies.

An Empirical Examination of Gender Stereotype from the Result of National Board Certification

The national board certification is an important initiative to recognize accomplished teachers across the United States. To date, more than 32,000 teachers in all 50 states have received the certification (Smith & Oliver, 2004). Other professional organizations, such as the National Council for Accreditation of Teacher Education (NCATE) and the Interstate New Teacher Assessment and Support Consortium (INTASC), are taking steps to align their accreditation processes with the national board standards (Goldhaber, Perry, & Anthony, 2003). As a result, the state licensing systems are designed to set minimum standards for novice teachers, and the National Board of Professional Teaching Standards (NBPTS) delineate what accomplished teachers should know and be able to do for advanced certification (Margolis, 2004).

Over the last 15 years, more than 500 school districts across the nation have implemented policies and regulations to recruit, reward, and retain teachers with the national board certifications (<http://www.nbpts.org/about/state.cfm>). Meanwhile, a critical question for education stakeholders is whether the national recognition truly identifies accomplished teachers in various subject areas (Thirunarayanan, 2004). According to a social-cognitive theory, raters usually have well-developed stereotypes of men and women (Bauer & Baltes, 2002). An analysis of empirical data from the state of North Carolina suggested that “male teachers are less likely to be certified [by the national board]” (Goldhaber et al., 2003, p. 2). Given the variation of applicants’ teaching experience, the purpose of this investigation is to examine gender differences in

achieving the certification among four subject areas. Findings from this large-scale data analysis can facilitate identification of confounding variables associated with the gender stereotypes, and thus, reconfirm or disconfirm validity of the national board certification across school disciplines.

Literature Review

Many researchers have identified a need for a validation study of the national board certification (Gitomer, 1997; Goldhabar, Perry, & Anthony, 2003; King, 1994; Kraft, 2001). Margolis (2004) contended that those accomplished teachers deserved a raise in an amount of “a couple of thousand dollars more.” Thirunarayanan (2004) projected that “if the spending continues at the current rate, the billion dollar mark will be surpassed within a few years.” One of the fundamental justifications to support the spending is to find indisputable evidence that those certified teachers are truly at the top of their profession (Smith & Oliver, 2004).

Bond, Smith, Bakers, and Hattie (2000) conducted a pilot study on validity of the national board certification. When examining instructional performance between certified teachers and their peers who applied and did not receive the certification, Bond et al. (2000) found significant differences in 11 out of 13 dimensions. The pilot study had a limited scope, dealing with a small sample of teachers within two certification areas. Podgursky (2001) noted that not all the dimensions had a clear focus on student performance. In addition, because the significant difference did not appear on all dimensions, more research is needed to reconfirm or disconfirm the existing findings on national board certification.

“Fundamentally, the National Board assessment system subscribes to the belief that products of classrooms, such as videotapes and student work, are powerful and valid forms of evidence for making claims about teaching practice” (Gitomer, 1997, p. 9). Because these authentic assessment measures must be graded or interpreted by human beings, validity of the final certification not only depends on quality of the portfolio documentation, but also hinges on potential stereotype influences from each scorer. To ensure reliability from the portfolio scoring, all certification cases are double checked to maintain consistency of the assessment (Pearlman, 2002). However, the consistency checking cannot completely eliminate gender-stereotype concerns on the grading system.

By definition, *gender stereotypes* are socially shared beliefs about the characteristics or attributes of men and women (Cleveland, Stockdale, & Murphy, 2000, p.466). “Many occupations are gender stereotyped” (Ottati & Lee, 2002, p. 230), and the teaching profession is traditionally believed to be an occupation for females (Beyer, 1999; Ehrenberg, Goldhaber, & Brewer, 1995). In the aforementioned pilot study, two subject areas investigated were *English language arts* and *middle childhood/generalist* (Bond et al., 2000). Because female roles have been well perceived in both child caring and language development, gender stereotypes, if they played a role, tended to skew similarly in favor of female participation.

On the other hand, “Science is still a domain dominated by males, both in industry and academia, and little has been done to change its practice, let alone to change its fundamental structure” (Letts, 1997, p. 3). This gender stereotype might lead the general public to assume more effective science teaching from male instructors. Although the

scorers were “highly accomplished teachers in their own right” (Gitmer, 1997, p. 7), the literature also suggested that gender has been a significant determinant of teachers’ qualitative evaluation (Ehrenberg, Goldhaber, & Brewer, 1995). When the biased view was shared by most scorers, reliability of the national board certification might remain relatively high, but the validity could still be low due to the systematic prejudice.

In fact, the national board certification heavily relied on scorers’ qualitative interpretation of the professional teaching standards (Gitomer, 1997). “Stereotypes learned through socialization may affect academic performance even if a person does not believe the stereotypes” (Walsh, Hickey, & Duffy, 1999, p. 221). To examine validity of the certification outcome, data from applicants up to 2004 have been analyzed in this study across four subject areas (Table 1). Because of the inclusion of both so-called feminine (e.g, language arts) and masculine (e.g., science) subjects, the analyses of certification data may help disentangle issues of stereotypic scoring between male and female applicants in national board certifications.

Insert Table 1 around here

Cleveland, Stockdale, and Murphy (2000) noted that “the differences between ratings of men and women may be a consequence of the raters’ social-cognitive processes rather than the sex of the rater” (p. 467). Consequently, scorers might share the same view about the gender role regardless of their gender identities. Instead of labeling scorers according to their gender identities, empirical data need to be analyzed to

disentangle contextual factors related to applicant evaluation during the national board certification.

Research Questions

In the existing setting of the national board assessment, an applicant's gender is known to each scorer from a review of the videotape documentation. In addition, the national board certification requires applicants having at least three-year teaching experiences (Kraft, 2001). Research questions that guide this investigation are:

1. Given the existence of gender and subject differences, are applicant scores linked significantly to the length of their teaching experiences?
2. What is the pattern of gender difference in receiving the national board certification among the four subjects?
3. Do score differences support the "certified" and "uncertified" decisions in light of the relationships between gender and subject categorizations?

Methods

National data from a total of 8279 applicants in four subject areas were provided by the National Board for Professional Teaching Standards to support a validation study of the certification outcomes. Applicants' *certified* vs. *uncertified* status along with their scores from the portfolio assessment were released in the database. In addition, the data also contain information on applicant gender, race, and number of years in the teaching profession.

In the previous pilot study, Bond et al. (2000) noted that teachers of other ethnicities did not apply for certification in large enough numbers to be effectively studied. Similarly, the race factor has been excluded from this investigation because

Asian, Hispanic, Black, and Native Americans altogether account for less than 12% of the applicant pool. In contrast, “One of the most salient groups, a basic category so to speak, is biological sex” (Beyer, 1999, p. 787). Jussim and Eccles (2002) concurred, “The finding that teachers seemed to be relying on sex stereotypes more than ethnic or social class stereotypes is broadly consistent with other research suggesting something uniquely powerful about sex stereotypes” (p. 267).

Whereas dummy variables can be employed to code categorical factors of gender and subject, the length of teaching experience was treated as a continuous variable to analyze its relationship with applicant scores. A regression analysis was conducted to accommodate these categorical and continuous factors, and the gender difference in applicant scores across subject areas are represented by an interaction effect between *gender* and *subject*. Since the applicant teaching experience is not gender specific, no interaction effects are needed between *teaching experience* and these categorical factors. Thus, Question 1 can be addressed by a regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 * X_2 + \beta_3 X_3 + e$$

Where Y = applicant score

X_1 and X_2 dummy variable and vector for gender and subject factors

X_3 is a continuous factor describing the length of teaching experience

β_i s accommodate regression coefficients

e is an error term from the regression analysis

“Essentially, the challenge for the [national board] assessment system is to make a single decision, to certify or not” (Gitomer, 1997, p. 10). Given the dichotomous outcome, odds ratio in the logistic regression is used to describe the likelihood of

receiving the national board certification (SAS Institute, 1990). Question 2 is answered by an examination of gender difference across subject areas. Whereas the national board certification is built on applicant scores from the portfolio and videotape assessments (Gitomer, 1997), the final decision also hinges on other qualitative outcomes. In the subject of reading, for instance, Kreft (2001) delineated,

In addition to these [scoring processes], candidates complete the Instructional Analysis Exercises, with an analysis of a beginning teacher's teaching at an assessment center. They also complete three two-hour essays on the teaching of literature, reading, and language development. (p. 13)

Question 3 can be resolved through a triangulation of the results between Questions 1 and 2 to cross-examine if the score difference really supported the *certified* and *uncertified* decisions in light of the relationships with the *gender* and *subject* categorizations.

Results

The applicant data clearly confirmed a stereotypic observation that teaching is a women's profession (Smulyan, 2004). Table 2 showed that females accounted for more than 85% of the overall applicant pool, and the dominance of female applicants appeared in every subject domain.

Insert Table 2 around here

Incorporating factors of gender, subject, and their interactions, Table 3 showed that the *year of experiences* had the largest *p* value, and was an insignificant factor behind the assessment scores.

Insert Table 3 around here

After deleting the *teaching experience* factor, all remaining factors were significant at $\alpha=.05$ (Table 4).

Insert Table 4 around here

Besides consideration of the portfolio assessment scores, the certification decision also hinges on applicants' completion of other qualitative tasks (Kreft, 2001). Switching the outcome measure from the portfolio scores to the dichotomous *certifying* or *non-certifying* decision, results from the logistic regression confirmed the insignificant role of the *teaching experience* factor (Table 5).

Insert Table 5 around here

Dropping out the *teaching experience* factor, Table 6 showed the linkage between the likelihood of certification and the *gender* and *subject* categorizations. More specifically, the gender effect was close to reaching the $\alpha = .05$ significance level (i.e., $p = .06$), and the *subject* and *gender*subject* influence remained significant after deleting main effect from the *gender* factor (Table 7).

Insert Tables 6 & 7 around here

Discussion

In comparison to other programs of teacher accreditation, establishment of the National Board of Professional Teaching Standards (NBPTS) in 1987 was a relatively new event. In 1954, the National Council for Accreditation of Teacher Education (NCATE) was established to develop standards for teacher education programs. Earlier than the NCATE was the American Association of Teachers Colleges established in 1927 to offer teacher accreditation in the U.S. (Kreft, 2001). Nonetheless, the NBPTS is unique in its claim to recognize accomplished teachers in the U.S. (Margolis, 2004).

As the NCATE and other agencies concurrently deal with the quality of teacher education, the NBPTS initiative still has the burden to prove its credibility of targeting at a higher level of the profession. For instance, Thirunarayanan's (2004) speculated that "the knowledge and skills expected of National Board Certified teachers is very much similar to the knowledge and skills required of beginning teachers." This hypothesis nullifies the professional difference, and, if being accepted, will inevitably lead to invalidation of the national board certification.

This investigation is focused on identifying contextual factors from both applicant and scorer perspectives that are closely linked to validity of the certification. From the applicant side, length of *teaching experience* is examined under the condition of *gender* and *subject* differentiations. Meanwhile, research literature on stereotypic perspectives has been reviewed to support an analysis of the interaction effect between *gender* and *subject* that is sensitive to scorers' opinions.

Length of Teaching Experiences

Teaching is a profession that takes time to mature and improve in various disciplines. Besides academic knowledge gained from academic programs, teachers need to learn various skills to facilitate student learning process. *Length of teaching experience* clearly reflects the fact that it takes time to polish these education skills in the teaching profession. On the other hand, research literature indicates a non-linear relationship between teaching experience and teacher effectiveness. In terms of problem solving in a classroom setting, Smith, Hall, and Woolcock-Henry (2000) noted that teachers with 11-20 years of experience were more optimistic overall regarding negative events than were those who had taught for over 20 years. Counter-examples similar to this observation suggest that it is not appropriate to assume that “the longer the teaching experience, the better the instructional outcome.”

Whereas no threshold of teaching experience is generally applicable to this seemingly curve-linear relationship, the national board requires applicants to have at least three years of teaching experience before applying for the advanced credential. Factoring in this regulation, the data analyses show that the length of teaching experience is no longer significant for achieving national board certification (Tables 3 & 5). In practice, this result appears to reflect the reality that all applicants are competing on the same ground regardless of the extra years of teaching beyond the minimum of three years.

Certification Outcomes

A dichotomous decision (*certified vs. uncertified*) is made for each applicant in the national board certification. The judgment is primarily grounded on individual scores from the portfolio grading. Regarding the gender difference, the results unanimously show a p value near .05 (Tables 4 & 6), regardless of the outcome differentiation between

continuous scores and dichotomous conclusions. On the other hand, the *subject* and *gender*subject* effects were significant at $\alpha=.05$ in the final analysis (Table 7).

The research community generally agrees that interpretation of statistical difference should be articulated with consideration of effect size, an index represented by the real value differences between contrast groups (see Thompson, 1998). To facilitate the result discussion, effect sizes for the *gender* contrast have been computed in each subject area (Table 8), and the results show a reverse of the gender difference in the portfolio scores between *science* and other subjects.

Insert Tables 8 around here

In part, this is because “Women have traditionally made up the vast majority of the teaching force” (Letts, 1997, p. 2), and thus, female applicants seem to fit most teaching fields. The subject of *science* represents an exception due to extensive roles played by male participants (Smulyan, 2004). Despite the effort of NBPTS to avoid stereotypic judgment among its scorers (Gitomer, 1997), results from the national data analyses clearly differentiated the gender difference across female and male subjects. Besides the scorer training, the scoring outcome also depends on applicant preparation. On balance, a fundamental measure to narrow the gender gap seems to hinge on improvement of teacher education programs, particularly in those subjects with gender-biased representations.

Whereas the portfolio scores provide rich information on a continuous scale, the final decision for certification is dichotomous (certified vs. uncertified). Fortunately,

consistent findings have been obtained from different statistical methods that fit the scale difference. For the continuous portfolio scores, dummy variables have been adopted in the regression analysis to code the *gender* and *subject* categories (Ott, 1993). On the other hand, a logistic regression was used to model the likelihood of certification from the dichotomous outcomes because “differences on the logistic scale are interpretable regardless of whether the data are sampled prospectively or retrospectively” (SAS Institute, 1990, p. 1072-1073). As more data are being gathered from the national board certification, results from this analysis can be reconfirmed prospectively by more statistical analysis in the future.

In summary, the national board certification is designed to recognize accomplished teachers in various subject areas. Idealistically, certification should be solely based on applicants’ outstanding performance primarily documented in their portfolios, regardless of their gender identities. Nevertheless, based on analyses of the national board data from North Carolina, Goldhaber et al. (2003) reported that male teachers were less likely to receive certification. In this study, portfolio scores from the national board certification are examined across the gender and subject categories. The results show that male applicants outperform female applicants in science, despite the typical view of teaching as a female occupation (Smulyan, 2004). In non-masculine subjects, such as English and social studies, female applicants consistently received higher scores. Whereas the subject-specific finding seems to be different from that reported by Goldhaber et al. (2003), this investigation is built on a larger and more recent database, and the empirical results are in line with the social-cognitive theory that projects variation of gender stereotypic views among different subject areas (Bauer &

Baltes, 2002). As the national board continues certifying more teachers in the profession, more studies are needed to reconfirm the subject-specific nature of gender inequity in the nation.

References

- Bauer, C. C., & Baltes, B. B. (2002). Reducing the effects of gender stereotypes on performance evaluations. *Sex Roles, 47*, 465-476.
- Beyer, S. (1999). The accuracy of academic gender stereotypes. *Sex Roles, 40*, 787-813.
- Bond, L., Smith, T. W., Baker, W. & Hattie, J. (2000). *The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study*. Greensboro, NC: Center for Educational Research and Evaluation.
- Cleveland, J. N., Stockdale, M., & Murphy, K. R. (2000). *Women and men in organizations: Sex and gender issues at work*. Mahwah, NJ: Erlbaum.
- Gitomer, D. H. (1997, March). *Challenges for scoring performance assessments in the NBPTS system*. Paper presented at the 1997 annual meeting of the American Educational Research Association (AERA), Chicago, IL.
- Goldhaber, D., Perry, D., & Anthony, E. (2003). *NBPTS Certification: Who applies and what factors are associated with success?* Washington, DC: Urban Institute (ERIC Document Reproduction Service No. ED 475841).
- Jussim, L. J. & Eccles, J. (2002). Are teacher expectations biased by students' gender, social class, or ethnicity? In Y. Lee, L. J. Jussim, and C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences*. Washington, DC: American Psychological Association.
- King, B. M. (1994). Locking ourselves in: National stands for the teaching profession. *Teaching and Teacher Education, 10*, 95-108.

- Kraft, N. P. (2001, April). *Standards in teacher education: A critical analysis of NCATE, INTASC, and NBPTS*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Letts, W. J. (1997, March). *Gender forms self images: The case of five male secondary science teachers*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Margolis, J. (2004). A response to "The National Board Hoax." *Teachers College Record*, <http://www.tcrecord.org>, Article ID Number: 11277.
- Ott, R. (1993). *An introduction to statistical methods and data analysis* (4th ed.). Belmont, CA: Duxbury.
- Ottati, V. & Lee, Y. (2002). Accuracy: A neglected component of stereotype research. In Y. Lee, L. J. Jussim, and C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences*. Washington, DC: American Psychological Association.
- Pearlman, M. (2002, January). The architecture of the NBPTS certification assessments. Paper presented at the NBPTS Invitational Research Conference, Chicago, IL.
- Podgursky, M. (2001). Should states subsidize national certification? *Education Week*, 20 (30), 38, 40-41.
- SAS Institute (1990). *SAS/STAT user's guide*. Cary, NC: Author.
- Smith, B. P., Hall, H. C., & Woolcock-Henry, C. (2000). The effects of gender and years of teaching experience on explanatory style of secondary vocational teachers. *Journal of Vocational Education Research*, 25, 21-33.
- Smith, T. & Oliver, J. S. (2004, April). *An examination of the relationship between depth of student understanding and teacher national board certification status*. Paper

presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Smulyan, L. (2004). Choosing to teach: Reflections on gender and social change.

Teachers College Record, 106, 513-543.

Thirunarayanan, M. O. (2004). National board certification for teachers: A billion dollar hoax. *Teacher College Record*, <http://www.tcrecord.org>, Article ID Number:

11266.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools*, 5 (2), 33-38.

Walsh, M., Hickey, C., & Duffy, J. (1999). Influence of item content and stereotype situation on gender differences in mathematical problem solving. *Sex Roles*, 41,

219-240.

Table 1

Subject areas involved in this study*

	Social Science	Science	English	Generalist
Middle childhood				4889
Early adolescence			1450	
Adolescence	916	1147		

* Numbers inside the table are the sample sizes involved in this study after data cleaning.

Table 2

Number of applicants across the gender and subject classifications

	Social Science	Science	English	Generalist
Male	409	413	78	302
Female	507	734	1372	4587

Table 3

Results incorporating gender, subject, gender*subject, and teaching experience

Source	df	Mean Square	F	p value
gender	1	4649.92	3.14	0.08
subject	3	32941.20	22.28	0.00
gender*subject	3	3169.36	2.14	0.09
experience	1	52.88	0.04	0.85

Table 4

Results incorporating gender, subject, and gender*subject

Source	df	Mean Square	F	p value
gender	1	6223.65	4.17	0.04
subject	3	23848.88	15.98	0.00
gender*subject	3	3916.65	2.62	0.04

Table 5

Logistic regression results incorporating gender, subject, gender*subject, and teaching experience

Effect	df	Wald Chi-Square	p value
gender	1	1.81	0.18
subject	3	93.13	0.00
gender*subject	3	9.20	0.03
experience	1	0.22	0.64

Table 6

Logistic regression results incorporating gender, subject, and gender*subject

Effect	df	Wald Chi-Square	p value
gender	1	3.39	0.06
subject	3	49.72	0.00
gender*subject	3	10.51	0.01

Table 7

Logistic regression results incorporating subject and gender*subject

Effect	df	Wald Chi-Square	p value
subject	3	74.25	0.00
gender*subject	3	10.18	0.02

Table 8

Mean Scores and Effect size of the gender effect

	Social Science	Science	English	Generalist
Male	284.16	290.93	296.36	279.26
Female	290.74	288.68	299.21	284.63
Effect Size	6.58	-2.26	2.85	5.38