Differential Item Functional Analysis by Gender and Race of the National Doctoral Program Survey

Benita J. Barnes and Craig S. Wells
University of Massachusetts, Amherst, MA, USA

barnesbj@educ.umass.edu; cswells@educ.umass.edu

Abstract

One way that policies get enacted in higher education is through educational research. In 2000 the National Association of Graduate-Professional Students conducted the National Doctoral Program Survey (NDPS) in an effort to learn more about doctoral students' experiences and to influence doctoral education policy at both the local and national level. However, the National Doctoral Program Survey (NDPS) have only been reported in the aggregate. This aggregate reporting is appropriate if the items on the survey are measuring the same construct with the same level of accuracy across all respondents, but if this is not the case, then the veracity of the study results can be severely compromised. The purpose of this study was to examine the NDPS instrument using differential item functioning (DIF) analysis to determine if survey items functioned differently across gender and race/ethnicity. We identified 29 of the 48 items as displaying DIF, meaning women and students of color were either more likely or less likely to agree with their Caucasian male peers on certain items. Therefore, some caution may need to be exercised when interpreting the NAGPS data for diverse groups of students.

Key Words: Doctoral education, DIF, Survey research, National Doctoral Program Survey, Gender, Students of color.

Introduction

Doctoral training in the United States is reputed to be one of the best systems of education in the world (Golde, 2006; Nettles & Millet, 2006). However, despite such accolades, graduate education in the U.S. has been plagued with some exceedingly vexing problems (Golde, 2006; Lovitts, 2001). Chief among them is high attrition rate (Bowen & Rudenstine, 1992) and the length of time it takes to complete a doctorate (Golde & Dore, 2001). The number of doctoral students who do not complete their doctorate has been described as "scandalously" high, particularly when compared to the completion rates of students who are pursuing professional degrees in areas such as law and medicine (Gravios, 2007). It has been estimated that between 40 and 50 percent of all students who enter doctoral programs do not finish (Bair & Haworth, 2004 Bowen & Rudstein, 1992; Golde, 1998, Nettles & Millett, 2006). Although we need not automatically assume that all attrition from doctoral programs is bad or spells failure for the student (Ehrenber, Zuckerman, Groen, & Brucker, 2009), the reasons for doctoral attrition, nevertheless, are poorly understood.

Editor: Nicole Buzzetto-More

This attrition is most often attributed to a lack of funding, poor advising and mentoring, and a lack of department and disciplinary integration (Lovitts, 2001). In addition to high attrition rates, the time it takes to earn a doctoral degree has also proven to be problematic. According to the findings for the Council of Graduate Schools' *PhD Completion Project,* it can take up to ten years for students to complete their doctorates (Gravios, 2007). Factors thought to contribute to time to degree include discipline, advising, department climate, and lack of financial support (Ferrer de Valero, 2001).

Given the documented concerns and challenges to graduate education, assessing the status and quality of graduate education in the United States has become a central concern for many entities—foundations, government agencies, businesses and industry, universities, accrediting agencies, and educational associations—over the last decade. Nowhere is this more evident than in the number of reports and national studies conducted in an attempt to understand various student experiences, outcomes of doctoral education, and ways in which doctoral education may need to be systematically reformed or re-envisioned (Golde & Dore, 2000; Lovitts, 2001; National Association of Graduate-Professional Students, 2001; Nettles & Millet, 2006; Nyquist & Woodford, 2000).

In 1997, Nettles and Millet (2006) conducted a national survey of doctoral students' experiences. Their study, entitled *Survey of Doctoral Student Finances, Experiences, and Achievement,* included 9,036 doctoral students from twenty-one institutions and eleven fields of study. The Nettles and Millet study examined nine key dimensions of doctoral education—financing; socialization; research productivity; satisfaction, performance, and progress; and rate of progress, completion, and time to degree. One of the major findings from this study suggested that significant gaps exist in the experiences of students of color and female graduate students compared to their White and male counterparts.

As a part of the *Re-envisioning the Ph.D.* project, Nyquist and Woodford (2000) assembled information from several hundred interviews, numerous focus groups, and over four hundred articles and other documents to identify the concerns of various stakeholders about doctoral education. The stakeholders included employers, higher education institutions, business and industry, government agencies, foundations, educational associations, and doctoral students. According to their monograph, doctoral students reported seven major concerns: 1) unclear expectations for faculty academic careers; 2) concern about the quality of faculty life; 3) the narrow definition of professional work; 4) the lack of quality mentoring and support from faculty; 5) disappointment with the direction provided by mentors; 6) threats to graduate funding; and 7) the desire to situate their learning in the context of the global economy.

A particular stream of research that has gained national attention because of concerns unearthed during the *Re-envisioning the Ph.D.* project focused on the purposes and outcomes of doctoral training. As a result, Golde and Dore (2001) conducted a national study that included 27 institutions, 11 arts and sciences disciplines, and 4,000 respondents. The intent of their study was to gain insight into how doctoral students perceive their graduate education, particularly as it related to the purpose, the content, and the various processes that are part and parcel to the doctoral school experience. Golde and Dore discovered that doctoral students are not trained for the jobs they want, nor are they trained for the jobs they eventually fill. More specifically, they are primarily trained to do academic research; however, most academic jobs are at liberal arts or comprehensive universities where the primary focus is on teaching rather than research. Secondly, Golde and Dore discovered that many students want more breadth from their doctoral experience particularly as it relates to taking courses outside of their discipline. Furthermore, according to Golde and Dore, students generally do not understand the process of doctoral education. Many students reported that they do not have a clear understanding of certain aspects of the doctoral degree process such as advisor expectations, time to degree, and how to obtain research funding.

In addition to the several large-scale studies that have been conducted in an effort to increase our understanding of various aspects of doctoral students' experiences, several reports have also gained national attention because of their focus on the need to improve the doctoral degree process. More specifically, The Committee on Science, Engineering, and Public Policy (COSEPUP) (1995) released a report that made two recommendations on how graduate programs, particularly at the department level, could implement several reforms to enhance the educational experience of doctoral students. The first recommendation was to offer a broader range of academic options. This recommendation called for departments to allow their students to gain a wider variety of skills so that they will be employable both inside and outside of the academy. The committee argued that skills that one needs for a career in the academy are different from the skills that one needs for a career in industry, for example. Their second recommendation was to provide better information and guidance to prospective and current graduate students. This recommendation called for departments to provide current students and potential students information about time to degree and employment options and opportunities so that they will be able to make informed employment preparation decision early in their training or even before they start their training.

Similarly, the Association of American Universities (AAU) (1998) recommended several ways in which departments could improve their doctoral training. Their recommendations included: 1) providing all admitted students with accurate information about the costs they will incur during their training; 2) balancing the breadth and depth of the curriculum to minimize time-to-degree; and 3) ensure more advisor-student interactions.

Using the previous large-scale studies as a springboard, but particularly influenced by the COSEPUP and the AAU reports, in 2000 the National Association of Graduate-Professional Students (NAGPS), which is a nonprofit, graduate student run and operated organization dedicated to improving the quality of life for graduate and professional students (Fagen & Wells, 2004), conducted the largest and most comprehensive doctoral student survey to date. This study was conducted in an effort to better understand graduate students' experiences and to assess students' perception of their programs' implementation of the educational practices that had been recommended by the COSEPUP and the AAU reports. The *2000 National Doctoral Program Survey* (NDPS) included questions in nine content areas: information for prospective students; curricular breadth and flexibility; teaching; professional development; career guidance and placement services; time to degree; faculty mentoring; program climate; and overall satisfaction. Key findings that were reported from the NAGPS' study demonstrated that 81 percent of the respondents were satisfied with their doctoral program, 86 percent were satisfied with their advisors, and 80 percent would recommend their program to prospective students (Fagan & Wells, 2004).

Although the results of the NAGPS' survey data may be useful for providing broad insights into the process and outcomes of doctoral education, the results have not been particularly useful for helping us understand how the process and outcomes of doctoral education may differ for different sub-populations of doctoral students. To date, the NAGPS' data have only been reported in the aggregate. Aggregate reporting is appropriate if the items are measuring the same construct such that the scores have the same meaning for all respondents (Dodeen & Johanson, 2003). However, if this is not the case, then the veracity of the study results can be compromised.

The purpose of this study is to examine the NAGPS' *2000 National Doctoral Program Survey* (NDPS) instrument to assess if the items are functioning differently for women and students of color. We elected to analyze the NAGPS data set for several reasons. First, it is numerically the largest national study conducted on doctoral students. Second, the study represented the greatest diversity with respect to the number of institutions and fields of study represented. Third, there was considerable overlap of topics covered on this survey and on the other two national studies. Finally, in addition to understanding doctoral students' perceptions of their program and their

experiences, the NAGPS survey was intended to impact doctoral education policy at the department, institutional, and/or national level.

Differential item functioning (DIF) was the analytical tool employed in this study because it allowed us to examine the instrument at the item level, which provides insight into whether women and students of color may be responding to items differently than their male and White counterparts. In the United States, there has been an increase in the number of women and students of color who are entering doctoral programs (Council of Graduate Schools, 2006; Hoffer et al., 2006). However, despite these increases, as Nettles and Millett (2006) identified, there is often a gap between the experiences that women and students of color have compared to their male and White counterparts. Therefore, it is becoming increasingly more necessary to examine survey instruments that are used to evaluate doctoral students' experiences to ensure that they are free of unintended influences based on the sex and race/ethnicity of the respondent. DIF can accomplish this purpose by comparing item statistics (e.g., difficulty and discrimination) among the respective groups, but controlling for overall level of experience in graduate school. DIF is described in more detail below.

Method

Participants

The *2000 National Doctoral Program Survey* is a 48-item summated rating scale based on nine content areas (see the Appendix for a list of the survey items). The survey was administered electronically from March to August of 2000. The NAGPS survey team widely publicized the survey to generate as broad a participation rate as possible. Major disciplinary societies, professional associations in each academic field, and other university personnel were encouraged to have their doctoral students complete the survey (Fagan & Wells, 2004).

Over 32,000 current or former doctoral students representing 5,000 doctoral programs at nearly 400 graduate institutions in the United States, including several historically Black colleges and Universities (HBCU), and Canada participated in the study. The demographic composition of the sample was as follows: 86% of the sample was white, 6 % Asian American, 4 % African American, 4 % Hispanic, and less than 1% American Indian. Women represented 51% of the sample. The institutions in the sample were classified as either research extensive, research intensive, or specialized. The disciplines included education, engineering, humanities, life science, physical science, social science, communication, and professional (Fagan & Wells, 2004).

Instrument

As noted above, the survey included items from nine content areas which are described below. For all items in the content areas a 4-point Likert-type scale was used: strongly disagree, disagree, agree, and strongly agree. Table 1 describes the nine content areas and lists the number of items per content area.

Table 1: Description of the content areas for NDPS instrument

Content Area	Description of Content Area	Number of Items
Information for Prospective Students	Assess how much information students had about their programs, such as the amount of time it takes to earn the degree and graduation rates, prior to them entering their doctoral programs.	8
Curricular Breadth and Flexibility	Assesses students' perceptions of their curricular options.	4
Professional Development	Assesses students' perceptions of being able to develop professional skills such as public speaking and grant writing.	3
Teaching	Assesses students' teaching experiences	4
Career Guidance and Placement Services	Assesses students' perceptions of the effectiveness of their career guidance.	2
Time to Degree Completion	Assesses students' perception of their degree progress.	3
Mentoring	Assesses students' relationship with his/her advisor.	7
Program Climate	Assesses students' perceptions of community and support within their programs.	7
Overall Satisfaction	Assesses students' satisfactions with such things as their program, advisor, courses, and standard of living.	6

Differential item functioning (DIF) analysis

Differential item functioning (DIF) occurs when a test or a survey item functions differently for a reference group (e.g., males) of examinees or respondents compared to a focal group (e.g., females) of examinees or respondents, after controlling for the level of the attribute being measured (Dodeen & Johanson, 2003; Kamata & Vaughn, 2004). For example, an item exhibits DIF if the probability of males responding to a specific category differs from females when they both are operating at the same overall level on the construct. Although DIF analyses are often conducted on achievement tests, assessing DIF in a non-cognitive assessment tool such as the *2000 National Doctoral Program Survey* instrument is important as it determines whether there are potential differences between respective groups after controlling for overall experience in graduate school as indicated by their perceptions/opinions. DIF analyses help determine the quality of each item on the instrument and whether the survey is measuring the construct in a similar manner in both groups. Furthermore, a DIF analysis increases the usefulness of using surveys as a tool of collecting data (Dodeen & Johanson, 2003).

There are several readily available statistics for detecting DIF in polytomous items (e.g., Likert-type). DIF statistics may be classified as either model-based or observed-score (non-model) based. One of the more popular model-based methods, due to its control of Type I error rate and acceptable power, is the likelihood ratio (LR) test (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinber, & Wainer, 1988). The LR test essentially compares the fit of a compact and augmented model to test for DIF between a reference and focal group (it is possible to test DIF among more than two groups simultaneously). The compact model constrains the item parameter values to be equal across the reference and focal groups (i.e., assumes no DIF is present). The augmented model allows the parameter values for one item (or a set of items) to be freely estimated in each group, constraining the remaining items to be equal across groups. DIF is assessed by comparing the overall fit of both models. If the item being tested contains DIF (i.e., the parameter values are not equal across groups), then the overall fit for the augmented model will be much better than

the overall fit for the compact model. The overall fit of the respective model is provided by -2 times log likelihood (-2 Log *L*). Because the compact model is hierarchically nested within the augmented model, the difference between the overall fit statistics ($[\text{-2 Log L}]_C - [\text{-2 Log L}]_A$) is distributed as a chi-square with the degrees of freedom equal to the number of parameters being tested, where C and A refer to the compact and augmented model, respectively.

An appropriate model used for Likert-type items is Samejima's (1969) graded response model (GRM). Samejima's GRM models the probability that an examinee responds to a particular category given her or his proficiency level, denoted θ. Samejima's GRM may be classified as a "difference" or "indirect" model in that the conditional probability of an examinee responding to a particular category requires a two-step process. The first step is to model the probability that an examinee's response falls in or above a particular category given θ. The probabilities, denoted $P_{ik}^{*}(\theta)$, may be computed as follows:

$$P_{ik}^{*} = \frac{\exp\left[a_i\left(\theta_j - b_{ik}\right)\right]}{1+\exp\left[a_i\left(\theta_j - b_{ik}\right)\right]}, \tag{1}$$

where $P_{ik}^{*}(\theta)$ referred to as operating or boundary characteristic curves, indicates the probability of scoring in the k^{th} or higher category on item *i* (by definition, the probability of responding in or above the lowest category is $P_{ik}^{*} = 1.0$); a_i refers to the discrimination for item *i*; b_{ik} refers to the threshold parameter (the threshold parameter is analogous to the item difficulty parameter for an achievement test).

Once the operating characteristic curves are estimated, the category response curves, which indicate the probability of responding to a particular category given θ, are computed by subtracting adjacent $P_{ik}^{*}(\theta)$ as follows:

$$P_{ik} = P_{ik}^{*}(\theta) - P_{i(k+1)}^{*}(\theta). \tag{2}$$

By definition, the probability of responding above the highest category is $P_{i(K+1)}^{*} = 0$; therefore, the probability of responding in the highest category is simply equal to the highest operating characteristic curve.

The equality of a set of item parameters within Samejima's GRM were tested via the LR test in the present study. The compact model constrained the item parameters (i.e., a_i and b_{ik}) to be equal between the reference and focal groups while the augmented model allowed the item parameter values for item *i* to be unconstrained (the remaining item parameter values were still constrained between the reference and focal groups). The LR test statistic, denoted G^2, is the difference in the overall fit in the two models, that is,

$$G^2 = -2\log L_C - \left(-2\log L_A\right). \tag{3}$$

G^2 is distributed as a chi-square with degrees of freedom equal to the difference in the number of parameters being estimated (or, in other words, the number of parameters being tested, which in this case is four – three thresholds parameters and one slope).

Factor analysis

The factor analysis was performed using polychoric correlation coefficients created by the software package PRELIS because the item responses were Likert-type (i.e., ordinal level). Before

conducting the DIF analyses, we wanted to evaluate the dimensionality of the instrument to confirm the data were appropriate for Samejima's GRM, which assumes the data are unidimensional. The factor analysis was performed using polychoric correlation coefficients created by the software package PRELIS because the item responses were Likert-type (i.e., ordinal level). The software package SPSS was used to perform a factor analysis to determine the number of underlying dimensions being measured by the survey. The sample of 32,676 was split in half. The first sample was used to identify the number of factors while the second sample was used for cross-validation purposes. The extraction method used was principal axis factoring. The number of underlying factors was determined via inspection of a scree plot and whether the percentage of variance explained was sufficient. Once the number of factors was determined, the rotated factor loadings were inspected to determine what each factor represented. An item was considered to be loading on a factor if the rotated factor loading was greater than 0.3. A value of 0.3 was used so that an item would need to exhibit a meaningful loading to be considered indicative of a particular factor and so that the majority of items would still be considered (i.e., if a larger criterion was used, it was possible that some useful items may not be included in the analysis). The rotated factor loadings were obtained via a promax rotation.

DIF analyses

Second, for each factor observed, the LR test was used to test each item for DIF for groups based on gender (female/male) and race (African American, Asian American, Caucasian, and Hispanic). It was important that the DIF was performed separately for each observed factor because the appropriate application of Samejima's GRM requires a unidimensional scale, which each factor provides. The software package IRTLRDIF (Thissen, 2001) was used to perform the LR test. Each item was tested at an overall significance level of 0.05. For each item identified as exhibiting DIF, we used the difference between the average adjusted item threshold parameter estimates between the respective groups to judge the meaningfulness of the DIF. Differences greater than 0.20 represented a meaningful magnitude of DIF.

Results

Dimensionality

Inspection of the scree plot of the factor analysis from the first sample (see Figure 1) revealed a strong first factor, which accounted for 34% of the variance, along with a weak second factor. The total variance explained by the two factors was 40%.

After inspecting the oblique rotated factor loadings, 40 items loaded on factor one (i.e., exhibited factor loadings greater than 0.30) while the remaining 8 items loaded on factor two. None of the items were cross-loaded on both factors. Most of the items that loaded on factor one dealt with the structure of the graduate school program while the items that loaded on factor two pertained to the relationship with the advisor. Therefore, we decided to refer to factor one as Program Structure while factor two was labeled Advisor Relationship. The correlation between factor one and factor two was 0.63, indicating that students who tended to respond positively with respect to their perception of the program structure also responded positively regarding their relationship with their advisor.

The final solution based on the factor analysis of the first sample was cross-validated using a second, independent sample. The results from the cross-validation using the second sample were the same as those based on the first sample.

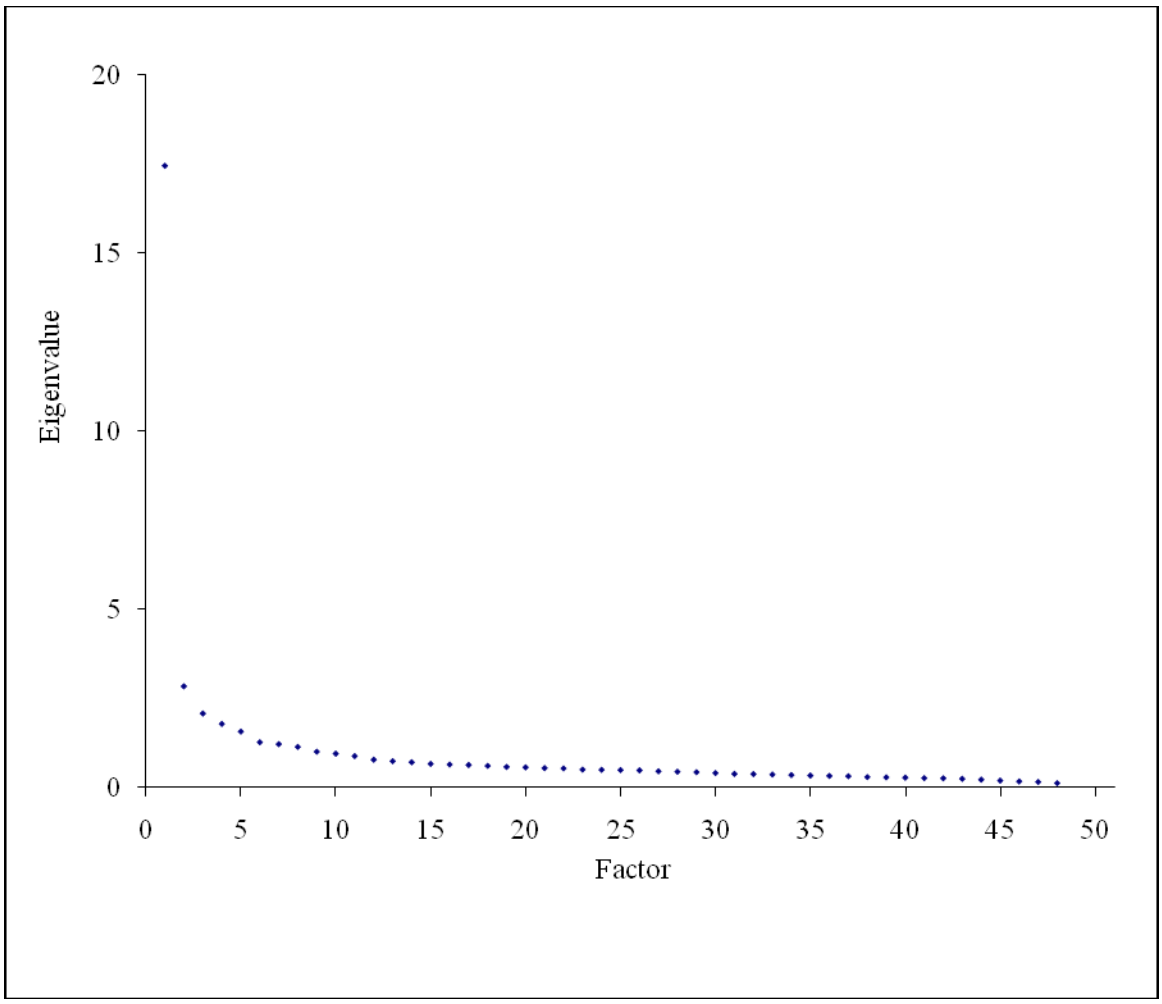

Figure 1: Scree plot indicating a dominant first-factor and a weak second factor

DIF analyses: Gender

The items within each factor were tested for DIF between females and males. Table 2 reports the DIF items along with the overall difficulty of the item (i.e., average threshold parameter estimate) for females and males. Larger overall item threshold values indicate that individuals from one group (e.g., females) were less likely to respond positively (i.e., agree) with the statement compared to individuals from the other group (e.g., males) even though the two groups were adjusted such that they were equivalent on their overall perception of their program structure.

Of the 40 items examined for the first factor, Program Structure, seven items displayed DIF. Compared to males, females were less likely to agree with the statement in four of the items (items 20, 42, 43 and 44), whereas they were more likely to agree with the statement in three of the items (items 18, 25 and 40). Three of the four items females were less likely to agree with than males centered on resources. In this case, resources included physical, financial, and time and they represent overall perceptions about resources for doctoral students collectively (i.e., doc students receive sufficient resources) as well as perceptions about individual resources (i.e., I receive sufficient financial support…, and I have enough free time…). The fourth item females were less likely to agree with was related to course satisfaction (i.e., I am satisfied with the courses in my program).

The three items that females were more likely to agree with compared to males did not follow a particular pattern but could be categorized as training, assessment, and recruitment. Two of the items assessed their perception at the program level (i.e., *Doctoral students* receive training in professional ethics and professorial responsibilities…, and *My program* actively recruits talented students from underrepresented groups). The third item assessed their experience at an individual level (My program *gives me* a clear, annual assessment of *my* progress).

Table 2: Average threshold parameter estimates for factor one, Program Structure, identified as functioning differentially between females and males

		Comparison	
Item	Question	Females	Males
18	Doctoral students in my program receive training in professional ethics and professorial responsibilities via coursework or seminars.	0.04	0.37
20	Doctoral students in my program receive sufficient resources such as office space, computer access, office equipment, and supplies.	-0.03	-0.40
25	My program gives me a clear, annual assessment of my progress towards the PhD.	-0.09	0.13
40	My program actively recruits talented students from underrepresented groups.	-0.63	-0.37
42	I have enough time and freedom to pursue interests outside of my academic program.	-0.04	-0.39
43	I receive sufficient financial support to maintain an acceptable standard of living.	0.44	-0.06
44	Overall, I am satisfied with the courses in my program.	0.07	-0.18

For the second factor, Advisor Relationships, none of the items were identified as displaying DIF based on gender.

DIF analyses: Race

The items within each factor were tested for DIF for each pairwise comparison for race: African-American versus Caucasian, Asian-American versus Caucasian, Hispanic versus Caucasian, African-American versus Asian-American, African-American versus Hispanic, and Asian-American versus Hispanic. The items flagged as exhibiting a meaningful magnitude of DIF for each comparison for factor one, Program Structure, and factor two, Advisor Relationship, respectively, are reported in Tables 3 and 4.

Table 3: Average threshold parameter estimates for DIF items for each pairwise comparison for factor one, Program Structure

	Comparison											
Item	Af-Am	Ca	As-Am	Ca	His	Ca	Af-Am	As-Am	Af-Am	His	As-Am	His
1			-2.11	-1.90			-1.82	-2.13				
2	-1.62	-1.39			-1.62	-1.41						
3												
4												
5												
6												
7	-0.09	-0.48	-0.15	-0.36	-0.11	-0.35						
8												
9			-1.19	-0.97								
10	-0.04	-0.26										
11	-0.22	-0.48	-0.26	-0.04			-0.05	-0.47			-0.05	-0.39
12	-1.56	-1.82	-1.66	-1.41			-1.36	-1.80			-1.36	-1.73
13	0.29	-0.08	0.30	-0.15	0.30	0.02						
14												
15	0.20	-0.20										
16												
17							-0.21	-0.54			-0.21	-0.53
18	0.05	-0.52	-0.05	-0.41	0.02	-0.32						
19					-0.25	-0.03	-0.35	-0.14			-0.36	-0.05
20							-0.78	-0.55				
21												
22	0.75	0.35	0.80	0.26	0.75	0.47					0.23	0.45
23												
24	0.82	0.46	0.86	0.33	0.83	0.47						
25												
26												
27			-0.78	-1.07			-1.05	-0.70			-1.05	-0.58
35	-1.39	-1.13										
36					-1.38	-1.10						
37												
38												
39			-0.94	-0.72	-0.92	-0.70						
40	-0.50	0.02	-0.44	-0.22			-0.23	0.01	-0.27	0.09		
41	-0.63	-0.11			-0.63	-0.39	-0.56	-0.11	-0.29	-0.02		
42			-0.04	-0.56			-0.83	-0.40			-0.55	-0.14
43			-0.40	-0.86								
44												
46												
47												
48												

Table 4: Average threshold parameter estimates for DIF items for pairwise comparison for factor two, Advisor Relationship

	Comparison											
Item	Af-Am	Ca	As-Am	Ca	His	Ca	Af-Am	As-Am	Af-Am	His	As-Am	His
28	-1.64	-1.31							-1.34	-1.14		
29												
30							-1.03	-0.82				
31												
32					-1.91	-2.38			-2.13	-1.54	-1.95	-2.36
33												
34	-2.55	-2.81	-2.61	-2.02			-1.95	-2.80			-1.96	-2.43
45					-1.32	-1.55						

African Americans versus Caucasians. Of the 40 items examined for the first factor, Program Structure, 13 items displayed meaningful DIF. African Americans were less likely to agree with nine of the items whereas they were more likely to agree with four of the statements compared to Caucasians. Two of the four items African Americans were more likely to agree with regarded the recruitment and environment for underrepresented groups ("My program actively recruits students from underrepresented groups." "My program provides an environment in which members of underrepresented groups feel comfortable and supported.").

Five of the nine items on factor one African Americans were less likely to agree with relative to Caucasians pertained to career preparation and guidance. In this case, career preparation included "exploring a broad range of career options," preparing students for "academic careers" and for "careers outside of academia." Career guidance included receiving effective "career guidance and planning services for careers outside of academia" and "placement assistance and job search support for positions outside of academia."

Of the eight items examined for the second factor, Advisor Relationships, two items displayed meaningful DIF. Interestingly, African Americans were more likely to agree with the item dealing with "learning good research practices" whereas they tended to respond less favorably relative to Caucasians to the statement "[m]y own goals and research interests are incorporated into my doctoral dissertation."

Asian Americans versus Caucasians. Fifteen of the 40 items examined for the first factor, Program Structure, displayed meaningful DIF. Asian Americans were more likely to agree with seven of the items compared to Caucasians. Similar to the African-American and Caucasian comparison, Asian Americans tended to respond more favorably to the statement pertaining to recruitment for underrepresented groups. Asian Americans were less likely to respond favorably to three statements regarding career preparation, guidance and placement. Asian Americans tended to respond less favorably relative to Caucasians regarding the statement, "[i]nsufficient funds slows my progress towards a degree."

Of the eight items examined for the second factor, Advising Relationships, one item displayed meaningful DIF. Unlike African Americans relative to Caucasians, Asian Americans were more likely to agree with the item dealing with incorporating personal goals and research interests into her/his dissertation relative to Caucasians.

Hispanics versus Caucasians. Of the 40 items examined for the first factor, 12 items displayed meaningful DIF. Hispanics were more likely to agree with six of the statements relative to Caucasians. Similar to the previous two comparisons, Hispanics tended to respond more favorably than

Caucasians regarding a comfortable supportive environment for members from underrepresented groups. In addition, Hispanics responded more favorably to the statement pertaining to the supportive community in the program. Three of the seven items on factor one Hispanics were less likely to agree with relative to Caucasians dealt with career preparation, guidance and placement, similar to the two previous comparisons.

Of the eight items examined for the second factor, Advisor Relationships, two of items displayed meaningful DIF in which both Hispanics were less likely to agree with relative to Caucasians. One of DIF items pertained to feeling comfortable discussing a career in academia with her/his advisor. The other statement Hispanics tended not to agree with as much as Caucasians dealt with overall satisfaction with her/his advisor.

African Americans versus Asian Americans. Of the 40 items examined for the first factor, Program Structure, 11 items displayed meaningful DIF. African Americans were more likely to agree with seven of the items. Interestingly, two of the seven items African Americans were more likely to agree with pertained to recruiting students from underrepresented groups as well as the supportive environment for members of underrepresented groups in the department. Asian Americans tended to respond more favorably to the statements regarding "preparing students for careers outside of academia" and that teaching experiences available through the program adequately prepares her/him "for an academic/teaching career."

Of the eight items examined for the second factor, two items displayed meaningful DIF. African Americans tended to respond more favorably regarding the amount of time s/he spends with her/his advisor whereas Asian Americans were more likely tended to agree with the statement pertaining to incorporating the student's goals and research interests into the dissertation.

African Americans versus Hispanics. Of the 40 items examined for the first factor, Program Structure, two items displayed a meaningful magnitude of DIF, both of which African Americans were more likely to agree with relative to Hispanics. The DIF items pertained to the recruitment of and environment for students from underrepresented groups ("My program actively recruits talented students from underrepresented groups." "My program provides an environment in which members of underrepresented groups feel comfortable and supported.").

Two items from factor two, Advisor Relationship, displayed meaningful DIF, both of which African Americans were more likely to agree with relative to Hispanics. The DIF items regarded "learning good research practices" and feeling comfortable talking to her/his advisor about a career in academia.

Asian Americans versus Hispanics. Six items for the first factor, Program Structure, were flagged as displaying a meaningful magnitude of DIF. Hispanics were more likely to agree with four of the items with relative to Asian Americans. Of interest, Hispanics were more likely to agree with statements pertaining to "career guidance and planning services for careers outside of academia." However, Asian Americans were more likely to agree with the statement, "[m]y program does a good job preparing students for careers outside of academia."

Of the eight items examined for the second factor, two items displayed meaningful DIF, both of which Asian Americans were more likely agree with relative to Hispanics. The DIF items pertained to talking with her/his advisor regarding careers in academia and that the dissertation incorporates the student's goals and research interests.

Discussion

Our DIF analysis of the NDPS revealed seven items had different response patterns between male and female respondents for factor 1—Program Structures. Three of the items females were more likely to agree with whereas the other four items females were more likely to disagree with com-

pared to males. One of the items that females more strongly agreed with was, "My program actively recruits talented students from underrepresented groups." One possible explanation as to why this difference occurred is female respondents might have been responding to the question very broadly (more underrepresented students are entering my program) whereas male respondents might have been responding to a specific word in the item such as "talented" (the underrepresented students they are recruiting are not talented). Although DIF alone cannot confirm this interpretation, it alerts us to the fact that there may be a potential problem with this item and that the wording might need to be examined before this question appears on future instruments. The second and third items that females responded more favorably to centered on being given clear assessment regarding their academic progress and their academic training as it pertains to professional ethics and professorial responsibilities, respectively. There is no clear explanation as to why different response patterns would emerge between females and males for this item. Therefore, this item bears further scrutiny.

Three of the four items that females were less likely to agree with pertained to "resources." For example, the first item addressed students' perceptions of students having sufficient physical resources such as office space, computer access, office equipment, and supplies. One plausible explanation as to why females did not respond as favorably to this item as their male counterparts is that male students still tend to have more or better assistantship (particularly research) opportunities than female students (Nettles & Millet, 2006); given this, male students may have greater access to resources such as office equipment and supplies. DIF in this case provides evidence, not of a flawed item, but of an important difference between females and males after controlling for overall program satisfaction. The second item centered on having sufficient funding to maintain an acceptable standard of living. Without access to additional data (e.g., information on the cost of their education or the amount of financial support they are receiving) there is no clear reason why female students would differ significantly in how they responded to this question than their male peers. However, the difference in responses does begin to raise an equity question with respect to whether or not female students are receiving less financial support than their male peers. The third item centered on time as a resource. Again, there are no intuitive or theoretical explanations as to why there would be differences in response patterns between female and male respondents on this item; however, questions do arise as to how female doctoral students might feel more overwhelmed and less able to balance their academic and personal lives. We found no evidence of gender DIF for factor 2—Advising Relationships.

Overall, we would recommend that each item that displayed DIF based on gender be looked at more carefully and, if necessary, be reworded. Some of the items that displayed DIF are double barreled questions—questions that ask the respondents for their perception of or opinion on two different things (Fowler, 1995). Item number 18, for example, which females were more likely to agree with, was double barreled because it asked about their academic training in both professional ethics and professorial responsibilities. If both male respondents and female respondents emphasized different parts of the question, DIF could result. It may be helpful to use a talk-aloud protocol to investigate the cognitive processes of both groups while responding to an item to understand the causes of DIF for these items.

Our DIF analysis, based on race, revealed that at least one item within each of the content areas displayed different response patterns among the various racial groups. When examining the items that displayed DIF between African Americans and Caucasians, Asian Americans and Caucasians and Hispanics and Caucasians, similar patterns emerged with respect to the items that Caucasians more favorably endorsed than their African American, Asian American and Hispanic peers. More specifically, there were four items that Caucasians responded more favorably to consistently than their peers from any of the underrepresented groups. These four items addressed the collective program experiences as they related to Curricular Breadth & Flexibility, and Career Guidance and

Placement Services. It would appear from the DIF analysis on these four items that Caucasians were more optimistic about the curricular options and the career preparation that their programs offer than their racially diverse peers. Although Caucasian students appear to be able to speak more broadly to the support their programs offer in these two content areas, this difference in response patterns may actually be due to the extent to which Caucasians may be getting better access to career guidance and placement services than their African American, Asian American, and Hispanic peers.

Two of the five items that African American and Hispanic students were more likely to agree with compared to Caucasian students centered on the experiences of underrepresented students. Given the content of these two items, it makes sense that African Americans and Hispanics might respond more favorably to them because they impact the under-represented students more directly than their Caucasian peers, whom would have to infer what underrepresented students might be experiencing. Another item that both African Americans and Hispanics responded more favorably to than their Caucasian peers centered on receiving a realistic assessment of financial support during the application and admission process. This difference in the response patterns between African Americans and Caucasians is consistent with findings from Nettles and Millett (2006) who found that African Americans (and other underrepresented groups) are increasingly being offered more financial support than Caucasians at the time of admissions.

When comparing items that displayed DIF between Asian Americans and Caucasians, seven items were identified in which Asian Americans tended to respond favorably from factor 1 and one item from factor 2. Although there were no distinct patterns to the individual items that displayed DIF from factor 1, several of the items centered on either Curricular Breadth & Flexibility or Campus Climate. The one item that displayed DIF from factor 2 pertained to respondents feeling like they had the opportunity to incorporate their own research interest into to their dissertation. There does not seem to be a theoretical justifiable reason why this difference would occur. Therefore, each item should be examined carefully to determine the item's usefulness and to determine if the items need to be reworded.

When underrepresented groups were compared with one another for DIF, several items continued to reappear. For example, the comparison between African American and Asian Americans identified eight items that African Americans responded more favorably to whereas Asian Americans tended to respond more favorably to five items. Similarly, when Asian Americans response patterns were compared to Hispanics, Asian Americans tended to respond more favorably on four items and Hispanics were more likely to agree with four items; however, all of the items had been found to favor one of the other groups in previous comparisons. The most interesting finding occurred between the African American and Hispanics. The DIF analysis between these two groups identified four items in which African Americans were more likely to agree with but found no items in which Hispanics were likely to agree with.

Limitations

The data analyzed in this study were drawn from an existing data set and, as such, the items were not necessarily designed for this study. However, the research design and analytic techniques employed in this research are wholly appropriate and desirable for secondary analysis of data. The data are also nine years old and may be less than ideal in terms of representativeness of current doctoral students; although, this data set is the best available source of survey data on doctoral students and continues to be analyzed and studied by other researchers. Additionally, it is worth noting that the limited categories for race used in these types of surveys often require that students choose a particular category (e.g. African American or Hispanic) when they may define their race/ethnicity differently. For example, some students may see themselves as bi- or multiracial, while others may identify with a more specific ethnic category (e.g., Puerto Rican or Chi-

cano(a)) within the Hispanic category. Finally, the original data set included only two HBCUs and therefore very few students from these institutions were included, making it difficult for any study that uses these data to contribute to our knowledge about doctoral education at these important but often under-resourced and under-studied institutions.

Conclusion

Of the 48 items on the NDPS, 29 of the items displayed DIF across different groups—gender or race. Fourteen of the items (1, 2, 9, 10, 15, 17, 20, 25, 28, 30, 36, 39, 44, & 45) displayed DIF once or twice while 15 of the items (7, 4, 12, 13, 18, 19, 22, 24, 27, 32, 34, 40, 41, 42, & 43) appeared at least three times and as many as five times. Two of the items (40 and 41) that displayed DIF across groups made substantive sense because those particular items spoke directly to respondents who belonged to a specific group. However, for most of the items there were no intuitive or theoretical explanations to justify the response differences. Despite the fact that more than half of the items on the survey displayed DIF, there is no single solution to resolve this problem. Removing the offending items on such a short survey could seriously jeopardize content representation, reliability, and construct validity. Nevertheless, the items that displayed DIF and for which there are no theoretical explanation need to be scrutinized further to determine how those items may be reworded in a way that makes all respondents equally likely to respond to the item the same way, after controlling for overall satisfaction. In particular, some of the double barreled items should be split to see if there are indeed differential patterns of response to each part of the double barreled questions. One solution may be to conduct some focus groups with doctoral students from different gender and racial/ethnic categories to ask them to explain how they understand the sources of difference for each item and to perhaps suggest rewording strategies that may strengthen these items in future surveys.

It is also worth noting that the DIF identified in some items may actually represent differences in the experiences of students from different groups rather than specific bias in the items. Those items in which the DIF can be explained by findings from previous studies or existing theory (e.g., items relating to access to resources as a source of difference between males and females) should be more closely investigated through further research into the differences between groups – for example, future studies might focus on gender inequities in terms of resources allocated to doctoral students. Given the findings of this study, future studies might also examine how advising about financial assistance is given to and interpreted by students from different racial/ethnic groups as they are selecting doctoral programs. The relatively large number of items with DIF related to racial and ethnic differences should caution us to be concerned about bias in this survey as well as on-going race/ethnicity-based inequities found in doctoral education. Future research may also want to examine these patterns in relation to other potential demographic differences (e.g., respondent's age, handicap, geographical location, major, employment, or stage of doctoral studies) among doctoral students.

The survey used in this study is also one of three national surveys that have been conducted on the experiences of doctoral students. It may be beneficial to spend more effort in building a more coherent survey that is better tested beforehand and then refined and replicated for continued use, rather than continually "re-inventing the wheel" with surveys that have not been as rigorously developed and tested. This type of strategy might yield an instrument that better captures the experiences of diverse students. This would also strengthen the validity and reliability of the instrument and, hence, enhance the rigor of the knowledge generated through such studies.

Survey instruments are an effective and an efficient way to learn about doctoral students' experiences and their perceptions of their doctoral programs. However, we need to continue to use sophisticated methods to advance our knowledge base about characteristics of items and different

population groups that allow items to behave differently. In the future, DIF needs to be assessed a priori when items are being constructed instead of after the survey has been disseminated.

Finally, although this paper has focused on the importance of developing survey items that will accurately represent students' perceptions regardless of gender or race/ethnicity, the overall findings from this study can provide useful implications for faculty members in doctoral programs. First, the findings from this study can be used by doctoral programs as an initial assessment tool to determine how they are doing in each of the nine content areas from the perspective of their female students and students of color. Second, findings from this study can be used to stimulate conversation among students and faculty regarding the graduate student experiences. Lastly, even when doctoral programs are demonstrating progress in their efforts to enroll female students and students of color, these programs should be wary of becoming complacent and assume that equity goals have been achieved. Women and students of color often times still feel they do not have access to the same resources as their male and White counterparts (Nettles & Millett, 2006) and doctoral programs need to be vigilant in their on-going efforts to address those issues in order to maintain and improve the ways in which they serve and retain these students.

References

Association of American Universities. (1998). *Committee on graduate education: Report and recommendations*. Washington, DC: Association of American Universities. Retrieved May 12, 2009, from http://www.aau.edu/reports/GradEdRpt.html

Bair, C. R, & Haworth, J. G. (2004). Doctoral student attrition and persistence: A meta-synthesis of research. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. XIX, pp. 481-534). Kluwer Academic Publishers, The Netherlands.

Bowen, W. G., & Rudenstine, N. L. (1992). *In pursuit of the Ph.D*. Princeton, NJ: Princeton University Press.

Committee on Science, Engineering, and Public Policy (COSEPUP) of the National Academy of Science, the National Academy of Engineering, and the Institute of Medicine .(1995). *Reshaping the graduate education of scientist and engineers*. Washington, DC: National Academy Press. Retrieved May 12, 2009, from http://www.nap.edu/readingroom/books/grad/sumary.html

Council of Graduate Schools. (2006). G*raduate enrollment and degrees 1986-2005*. Washington, DC: CGS.

Dodeen, H., & Johanson, G. A. (2003). An analysis of sex-related differential item functioning in attitude assessment. *Assessment & Evaluation in Higher Education, 28*(2), 129-134.

Ehrenberg, R. G., Zuckerman, J. A., Groen, J. A., & Brucker, S. M. (2009). Changing the education of scholars: An introduction to the Andrew W. Mellon foundation's graduate education initiative. In R. G. Ehrenberg and C. V. Kuh (Eds.), *Doctoral education and the faculty of the future* (pp.15-34). Ithaca, NY: Cornell University Press.

Fagen, A. P., & Wells, K. M. S. (2004). The 2000 National Doctoral Program Survey. In D. Wulff & A. E. Austin (Eds.), *Paths to the professoriate: Strategies for enriching the preparation of future faculty* (pp.74-91). San Francisco, CA: Jossey-Bass.

Ferrer de Valero, Y. (2001). Departmental factors affecting time-to-degree and completion rates of doctoral students at one land-grant research university. *Journal of Higher Education, 72*(3), 341-367.

Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation*. *Applied Social Research Methods Series, Volume 38*. Thousand Oaks, CA: Sage Publication.

Golde, C. M. (1998). Beginning graduate school: Explaining first-year doctoral attrition. In M. S. Anderson (Ed.), *The experience of being in graduate school: An exploration*. *New Directions for Higher Education, No.101* (pp.55-64). San Francisco: Jossey-Bass.

Golde, C. M. (2006). Preparing stewards of the discipline. In C. Golde & G. Walker (Eds.), *Envisioning the future of doctoral education: Preparing stewards of the disciplines* (pp. 3-20). San Francisco, CA: Jossey-Bass.

Golde, C. M., & Dore, T. M. (2001). *At cross purposes: What the experiences of doctoral students reveal about doctoral education*. Philadelphia, PA: A report prepared for The Pew Charitable Trusts. Available at www.phd-survey.org

Gravois, J. (2007). In humanities, 10 years may not be enough to get a PhD. *Chronicle of Higher Education, 53*(47), A1.

Hoffer, T. B., Welch, V., Jr., Webber, K., Williams, K., Lisek, B., Hess, M., Loew, D., & Guzman-Barron, I. (2006). *Doctorate recipients from United States universities: Summary report 2005.* Chicago: National Opinion Research Center. Retrieved May 12, 2009, from http://ww.norc.uchicago.edu/issues/docdata.htm

Kamata, A., & Vaughn, B. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal, 2*(2), 49-69.

Lovitts, B. E. (2001). *Leaving the ivory tower: The causes and consequences of departure from doctoral study*. Lanham, MD: Rowman & Littlefield.

National Association of Graduate and Professional Students. (2001). *Preliminary executive summary of the National Doctoral Program Survey*. Washington, DC: Retrieved May 12, 2009, from http://survey.nagps.org

Nettles, M. T., & Millett, C. M. (2006). *Three magic letters: Getting to Ph.D.* Baltimore, MD: The Johns Hopkins University Press.

Nyquist, J. D., & Woodford, B. J. (2000). *Re-envisioning the Ph.D. What concerns do we have?* Retrieved May 3, 2009, from http://www.grad.washington.edu/envision/project_resources/concerns.html

Thissen, D. (2001). IRTLRDIF v.2.0b [Computer program]. University of North Carolina at Chapel Hill: L. L. Thurstone Psychometric Laboratory.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). User of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Appendix

Item	Content[a]	Item Stem
1	IPS	My doctoral program provided me with accurate information about the cost of the program during the application and admissions process.
2	IPS	My doctoral program proved me with a realistic assessment of financial support while in graduate school during my application and admissions process.
3	IPS	My doctoral program provided me clear information about the requirements and expectations of my program during my application and admissions process.
4	IPS	My doctoral program provided me with information about career prospects for PhDs in my field during my application and admissions process.
5	IPS	My doctoral program provided me with a list of places where recent program graduates were employed after graduation during my application and admission process.

Item	Content[a]	Item Stem
6	IPS	My doctoral program provided me the percentage of students in the program who complete the program with a PhD during my application and admissions process.
7	IPS	My doctoral program provided me the average time to degree for recent program graduates during my application and admissions process.
8	IPS	Overall, my program provided enough information during the application and admissions process for me to make an informed decision about choosing to pursue a PhD.
9	CBF	My program's curriculum is broad enough to meet my needs and prepare me for my career choice.
10	CBF	My program encourages students to explore a broad range of career options.
11	CBF	My program encourages students to broaden their education through non-required activities.
12	CBF	My program does a good job of preparing students for academic careers.
13	CBF	My program does a good job of preparing students for careers outside of academia.
14	TE	Teaching assistants in my program are appropriately prepare and trained before entering the classroom.
15	T	Teaching assistants in my program are appropriately supervised to help improve their teaching skills.
16	T	Doctoral students needs and interests are given appropriate consideration for determining which courses students in my program teach.
17	T	The teaching experience available through my program is adequate preparation for an academic/teaching career.
18	PD	Doctoral students in my program receive training in professional ethics and professorial responsibilities via coursework or seminars.
19	PD	Doctoral students in my program receive training in professional skills such as pubic speaking, grant writing, and working in teams.
20	PD	Doctoral students in my program receive sufficient resources such as office space, computer access, office equipment, and supplies.
21	CGPS	Doctoral students in my program receive effective career guidance and planning services for careers in academia.
22	CGPS	Doctoral students in my program receive effective career guidance and planning services for careers outside of academia.
23	CGPS	Doctoral students in my program receive effective placement assistance and job search support for positions in academic.
24	CGPS	Doctoral students in my program receive effective placement assistance and job search support for positions outside of academia.
25	TDC	My program gives me a clear, annual assessment of my progress towards the PhD.
26	TDC	A group of faculty members (in addition to my advisor) is keeping track of my research progress and will help to determine when I have accomplished enough work for my PhD Degree.

Item	Content[a]	Item Stem
27	TDC	Insufficient funding slows my progress towards a degree.
28	M	I am learning good research practices.
29	M	I am receiving ongoing feedback regarding my Ph.D. progress.
30	M	I am satisfied with the amount of time I spend with my advisor.
31	M	I am satisfied with the quality of time I spend with my advisor.
32	M	I would feel comfortable talking to my advisor about a career in academia.
33	M	I would feel comfortable talking to my advisor about a career outside of academia.
34	M	My own goals and research interest are incorporated into my doctoral dissertation.
35	M	There is a person or office I would turn to if I perceived abuse or misconduct in my program by my advisor, or by a committee member.
36	PC	There is a supportive community in my program.
37	PC	Doctoral students in my program are treated with respect.
38	PC	Doctoral students in my program are involved in decisions relevant to their education.
39	PC	Faculty in my program believe students are here primarily to help faculty fulfill their research and teaching obligations.
40	PC	My program actively recruits talented students from underrepresented groups.
41	PC	My program provides an environment in which members of underrepresented groups feel comfortable and supported.
42	PC	I have enough time and freedom to pursue interests outside of my academic program.
43	OS	I receive sufficient financial support to maintain an acceptable standard of living.
44	OS	Overall, I am satisfied with the courses in my program.
45	OS	Overall, I am satisfied with my advisor.
46	OS	Overall, I am satisfied with my program.
47	OS	Overall, students in my program seem satisfied with the program.
48	OS	Overall, I would recommend my program to prospective students.

[a] IPS= Information for Perspective Students, CBF= Curricular Breadth & Flexibility, T= Teaching, PD= Professional Development; CGPS= Career Guidance & Placement Services; TTDC= Time to Degree Completion, M= Mentoring, PC= Program Climate, OS= Overall Satisfaction

Biographies

Benita J. Barnes, Ph.D., is an assistant professor at the University of Massachusetts Amherst in the department of Educational Policy, Research, and Administration. She holds her doctorate in Higher, Adult, and Lifelong Education from Michigan State University where she also earned a master's degree in Measurement and Quantitative Methods; a Master's degree in Adult and Continuing Education; and a B.A. degree in Psychology. Dr. Barnes's research includes graduate education, doctoral advising, identity development, and millennial students. Dr. Barnes is currently serving as concentration coordinator for the higher education administration program at UMass Amherst and is the past president (2003-04) of the National Association of Graduate and Professional Students.

Craig S. Wells, Ph.D., is an Assistant Professor of Education in the Research and Evaluation Methods Program in the School of Education at the University of Massachusetts Amherst. He received a doctorate and Master's degree from the University of Wisconsin – Madison in Educational Psychology with an emphasis in Quantitative Methods. Dr. Wells' research interests center on the development and application of quantitative methods in the behavioral sciences. He is particularly interested in the area of item response theory, especially as related to differential item functioning, assessing model fit, and modeling item response functions nonparametrically, and more generally in the areas of nonparametric statistics and the theory of hypothesis testing.