One-sided $T^2$ test for assessing the need for an Affirmative Action plan: A reanalysis of the *Shea v. Kerry*

Qing Pan *, Weiwen Miao † and Joseph L. Gastwirth ‡

August 19, 2019

KEY WORDS: Affirmative Action plan; Effect size; Employment discrimination; Under-representation of minority groups; One-sided Hotellings $T^2$ test; Required sample size.

Abstract

In the 1980’s, reports from Congress and the Government Accountability Office (GAO) presented statistical evidence showing that employees in the Foreign Service were overwhelmingly White male, especially in the higher positions. To remedy this historical discrimination, the State Department instituted an affirmative action plan during 1990-1992 that allowed females and race-ethnic minorities to apply directly for mid-level positions. A White male employee claimed that he had been disadvantaged by the plan. The appellate court unanimously held that the manifest statistical imbalance supported the Department's instituting the plan. One judge identified two statistical issues in the analysis of the data that neither party brought up. This paper provides an empirical guideline for sample size and a one-sided Hotellings $T^2$ test to answer these problems. First, an approximate rule is developed for the minimum number of expected minority appointments needed for a meaningful statistical analysis of under-representation. To avoid the multiple comparison issue when several protected groups are involved, a modification of Hotellings $T^2$ test is developed for testing the null hypothesis of fair representation against a one-sided alternative of under-representation in at least one minority group. The test yields p-values less than 1 in 10,000 indicating that minorities were substantially under-represented. Excluding secretarial and clerical jobs led to even larger disparities.

1. INTRODUCTION

In order to remedy a manifest imbalance in a traditionally segregated job category, an employer can establish an affirmative action plan (AAP). The US Supreme Courts
decision in Weber\(^1\) specified the criteria justifying an AAP. First, there must be a manifest imbalance (under-representation) in a traditionally segregated job category. Substantial statistical disparities are needed to show a meaningful imbalance and they should compare an employers work force to the labor force possessing the relevant skills and qualifications for the jobs in question.\(^2\) Second, the rights of non-minority employees should not be unnecessarily trammeled meaning the plan neither requires the termination of such employees so minorities could replace them or creates an absolute bar to their advancement. Finally, the preferences are temporary in duration.

Although the Court has not defined manifest imbalance, usually lower courts require statistically significant disparities with a meaningful effect size\(^3\). As the Shea v. Kerry decision necessarily relies in legal precedents using conventional hypothesis testing, the procedures and analyses presented here will do the same. Because statistical significance depends on the sample size and demographic mix of the appropriate labor force, the odds ratio measure of effect size will also be reported as this will aid courts in deciding when a disparity is substantial. In their review of effect sizes for 2x2 tables, Olivier and Bell (2013) suggest that the odds ratios of .82, .54 and .33 be used as cut-offs for small, medium and large effect. While these cut-offs are somewhat arbitrary, they aid in assessing the impact of a disparity and suggest that a small disparity may not correspond to a manifest imbalance, while medium and large ones do\(^4\).

In the mid and late 1980s both Congress and the GAO issued reports based on statistical evidence showing that employees in the Foreign Service, especially at the higher levels, were predominantly White and male. In contrast, minorities, especially females, were overwhelmingly concentrated in lower level clerical and secretarial positions. This can be seen from Table 1, which reports the race-gender composition of administrative jobs in the 1980, 1990 Censuses, and the demographic composition of Foreign Service Generalists in 1989 and the odds ratios for the State Department 1989 employees assuming their fractions of the qualified labor pool equals their fractions of administrative jobs in the 1990 census. To remedy these stark imbalances\(^5\), the State Department instituted an AAP for the 1990-1992 period that allowed females and race-ethnic minorities to apply directly for mid-level positions, without previously serving in an entry-level one.

In 2001, Mr. William Shea, a White male hired during the 1990-92 period, filed


\(^2\)The Johnson decision, Ibid. cites Hazelwood School District v. United States, 433 U. S. 299 (1977) (must compare percentage of Blacks in employer’s work force with percentage of qualified Black teachers in the area to determine whether they are underrepresented in teaching positions).

\(^3\)In Kohlbek et al. v. City of Omaha, 447 F.3d 552 (8th Cir. 2006) the court noted that a disparity between observed and expected hiring levels does not necessarily demonstrate possible discrimination, rather the numbers must be statistically significantly different.

\(^4\)The appropriate effect size measure and corresponding thresholds should be set by the legal system and may vary with the type of case and number of individuals affected. The authors believe that if the odds of a minority applicant being successful are one half those of a majority member the imbalance should be considered substantial.

\(^5\)Although the effect size measures were not in the record, it is noteworthy that, with the exception of Asian Americans, none of the odds ratios are small and most are in the medium and large category.
Table 1: Race-Gender Composition of individuals in administrative jobs in the 1980 & 1990 Census and the corresponding composition of Foreign Service Generalists in 1989 (OR in the table stands for odds ratio.)

<table>
<thead>
<tr>
<th>Year</th>
<th>White</th>
<th></th>
<th></th>
<th>Black</th>
<th></th>
<th></th>
<th>Hispanic</th>
<th></th>
<th></th>
<th>Asian American</th>
<th></th>
<th>American Indian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td></td>
<td>Male</td>
<td>Female</td>
<td></td>
<td>Male</td>
<td>Female</td>
<td></td>
<td>Male</td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td>1980</td>
<td>0.581</td>
<td>0.282</td>
<td>0.045</td>
<td>0.037</td>
<td>0.022</td>
<td>0.011</td>
<td>0.009</td>
<td>0.004</td>
<td>0.006</td>
<td>0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990</td>
<td>0.454</td>
<td>0.351</td>
<td>0.050</td>
<td>0.071</td>
<td>0.024</td>
<td>0.021</td>
<td>0.011</td>
<td>0.008</td>
<td>0.006</td>
<td>0.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1989</td>
<td>0.680</td>
<td>0.195</td>
<td>0.035</td>
<td>0.020</td>
<td>0.031</td>
<td>0.008</td>
<td>0.0016</td>
<td>0.008</td>
<td>0.003</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR</td>
<td>2.560</td>
<td>0.448</td>
<td>0.689</td>
<td>0.267</td>
<td>1.301</td>
<td>0.376</td>
<td>1.462</td>
<td>1.000</td>
<td>0.498</td>
<td>0.199</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: The race-gender distributions of administrative jobs from the 1990 national census was downloaded from https://www.census.gov/library/publications/1992/dec/cp-s-1-1.html. The information on Foreign Service employees is from the AAP; the table is on page 210 of Vol. 1 of the Joint Appendix in the appellate court record.

a complaint asserting that he started at a lower pay grade because of the AAP and that the plan infringed on his rights under Title VII of the Civil Rights Act. The courts found that he was qualified for a higher level appointment, but concluded that the AAP in the State Department was appropriate. In his concurring opinion, Judge Williams noted that neither partys submissions described the use of a statistical test or a standard of statistical significance. Furthermore, he noted that no adjustment was made for the multitude of comparisons when a variety of job groups were examined and questioned whether one can find statistical significance when the expected number of minority employers was less than one. Further background on the Shea vs. Kerry case is given in Appendix I.

The paper is organized as follows: Section 2 provides further information about the case, the decision and the statistical questions raised in the concurring opinion of Judge Williams. Section 3 modifies Hotellings $T^2$ test to test the null hypothesis of fair hiring against a one-sided alternative of under-representation following the ideas in Chernoff (1954) and extended by Self and Liang (1987). Section 4 describes the application of the test to data from the case. The modified $T^2$ test yields p-values less than 1 in 10,000 indicating that females and African-Americans were substantially under-represented in the major job categories. Excluding secretarial and clerical jobs led to an even larger disparities and more significant evidence supporting the AAP. After adjusting for multiple comparisons, the p-values remain less than 1 in 1000. These results confirm the courts decision that the imbalance of the female and African-American proportions between the State Department and the national labor force was substantial. Thus, the State Department was justified in implementing its AAP.

2. STATISTICAL ISSUES RAISED IN THE CONCURRING OPINION OF JUDGE WILLIAMS

Judge Williams agreed that the court had properly applied the two key Supreme Court cases, Johnson and Weber. He wrote separately to point out the vagueness of the term manifest imbalance and that two important statistical issues were not raised in the case. Our focus is on the statistical problems raised in the concurrence.

First, Judge Williams observed that the State Department's plan does not mention any statistical tests or significance level used to examine the data. Indeed, it is unclear that any statistical inferential analysis was carried out, i.e., the plan simply consisted of numerical comparisons of the demographic composition of various subgroups of employees to that of the national civilian labor force of public administrators and officials. A related problem was the lack of any adjustment in the standard for statistical significance to account for the multiplicity of subgroups, as would be necessary if we assume that State was seeking to identify only imbalances not attributable to random chance.

A second problem that concerned the judge was the relatively small number of employees in some of the subgroups that led to a complete lack of intelligible criteria for State’s assertions of manifest imbalance. In particular, he noted that the American Indian females formed only 0.2% of the available labor force for positions in the Finance Officer division of the State Department, which employs only 125 people. The concurrence continues with It seems improbable that any statistical test or standard of significance could yield evidence of a non-random imbalance for such a small subgroup.

3. STATISTICAL TECHNIQUES THAT ALLEVIATE THE PROBLEMS NOTED BY JUDGE WILLIAMS

3.1. Issues with Small Categories

In this section, the data for American Indian females in administrative positions will be examined. Then some guidelines for the minimum expected number of minorities needed for statistical analysis will be given.

Assuming the employees of the State Department were a random sample of the relevant subset of the national labor force identified by the EEOC, the expected number of American Indian females among the 125, is only .25 (125*0.2%) or one-fourth of a person. The probability of observing no American Indian females in the Finance Office Division is .7786, or just under 80%. This probability greatly exceeds the

---

7 796 F. 3d 42 at 66 (noting that the study in the record does not describe which, if any, statistical test was or standard of significance was used).
8 Ibid. at 66.
9 Ibid. at 66.
10 The number of American Indian females in the Finance Office Division follows a binominal distribution.
usual .05 level for determining statistical significance.\textsuperscript{11} Thus, it is mathematically impossible to infer a statistically significant shortfall of American-Indian Finance officers, substantiating Judge Williams’s argument that it seems improbable that any statistical test could yield statistically significant evidence of a shortfall in such a small subgroup, where hiring one minority would over-represent them.

Because there are many small minority sub-populations in the nation, the relationship between their fraction of the available labor force ($\pi$), the number of positions available ($n$) and the probability of observing or failing to observe a statistically significant shortfall in the number of minority hires ($x$) needs clarification. Two criteria will be described. The first answers the question raised by Judge Williams; what are the conditions $\pi$ and $n$ need to satisfy in order that observing 0 hires will be statistically significant? Alternatively, one can ask when is the probability a random sample of $n$ from the available labor pool will contain no minorities sufficiently large that one cannot obtain a statistically significant result, i.e. when it is mathematically impossible to find a statistically significant shortfall in minority hires.\textsuperscript{12} A related question concerns the power of the test, i.e. the probability of a statistically significant result when the odds a qualified minority member is hired are only one-half those of a majority? When will this probability be at least 0.5 or 0.8? When $n$ and $\pi$ have values enabling the proper statistical test to meet these conditions, finding a non-significant result is more meaningful and might support an inference that the defendant is fair.\textsuperscript{13}

When the minorities form a fraction, $\pi$, of the qualified and available labor force (QUALF), and the employer has hired (or employs) a total of $n$, courts assume that the employees of a firm that hires fairly are a random sample from the QUALF. Hence, the number $x$, of minority hires should be close to $n\pi$, their expected number in such a random sample. In statistical parlance, courts assume the number of minorities hired follows a binomial distribution with parameters $n$ and $\pi$. The Court originally adopted this model for the analysis of data pertaining to the representativeness of juries.\textsuperscript{14} For our purposes, the usual Poisson approximation to the binomial distribution will be used when the minority fraction, $\pi$, of the QUALF is small.\textsuperscript{15} Then, the expected number $n\pi$ of minority hires is the parameter of the Poisson distribution. Thus, the probability of observing no minority hires or employees equals $e^{-n\pi}$. For the data in Shea v. Kerry, Judge Williams considered $n = 125$, $\pi = .002$, so $n\pi = .25$ and

\begin{footnotesize}

\begin{enumerate}
\item[11]When statistical tests are used to check whether the data satisfies some assumptions, the .10 level is often adopted, see Fleiss (1986) and Cheng (2015). Gastwirth et al. (2009) found that the .15 level was appropriate when Levene’s test for homogeneity of the variance of several groups was used as a preliminary check. Of course, .78 is much larger than these levels of significance.
\item[12]Clearly, courts should not give any weight to a non-significant finding in this situation.
\item[13]The courts, not statisticians, should determine the value of the odds ratio or selection rates of the success of the minority and majority applicants that is legally important and the power or probability of obtaining a statistically significant result when the success rates differ by that value.
\item[14]For the binomial model, one standard deviation equals $\sqrt{n\pi(1-\pi)}$. Two-sided tests are used in the jury discrimination context because members of either gender or of any race-ethnic group could be subject to unfair treatment. The court accepted this test in Castenada v. Partida, 430 U.S. 482 (1977). The analysis of data from this case is described in Zeisel and Kaye (1997) and Finkelstein and Levin (2001).
\item[15]See, e.g. Theorem 4.2.1 in Larsen and Marx (2017).
\end{enumerate}

\end{footnotesize}
the probability of observing no minorities is $e^{-0.25} = 0.7788$ or nearly 80%. Thus, it is mathematically impossible for the appropriate statistical test to yield a significant result at the usual 0.05 level. One can ask how large must the sample size $n$ be in order for observing no minority employees in a position to reach statistical significance. Using the 0.05 level criterion, one needs $e^{-n\pi} = 0.05$, or $n\pi = \ln(20) = 2.9957$, or 3. Thus, if the expected number of minority employees in a job category is less than three, statistical significance cannot occur. This implies that when minority availability is 0.002, one needs a sample of size $n = 1500$ in order to conclude that an employer who has hired no member of the group has a statistically significant shortfall. Thus, when zero minorities occur in a small data set, such as the 125 Finance Officers noted by Judge Williams, formal statistical analysis of the data in that position is uninformative.

To illustrate the usefulness of the expected number, we examine the 1989 data on American Indian Female employees in three other positions for Foreign Service specialists. There were 4964 Foreign Service Generalists, of whom five were American Indian Females and their availability was 0.3%. Thus, one expects 14.89 (4964*0.003), or nearly 15. The probability of observing five or fewer members of this minority in a sample of 4964 is only 0.003. Thus, there was a statistically significant shortfall of American Indian females, indicating under-representation among the State Departments Foreign Service generalists. In other job categories there were very small proportions of female American-Indians in the QUALF, so the expected numbers of hires were less than three and statistical tests are meaningless.

While an expected number of three is a minimum number, it does not provide adequate power to detect a meaningful disparity. For example, in 1989, there were 1983 administrative employees, only one of whom was a female American Indian. Given their availability for those jobs equaled 0.2%, one expects 3.996, virtually four. Had there been no American Indian females, the data would be significant (p-value=.0188). Because the probability of observing one or fewer minorities in a sample of 1983 from this labor pool is 0.94, the data would not be significant at the usual 0.05 level. However, the power of the test is very low. The American Indian female data raise two important questions:

---

16 The binomial calculation also gives 0.7788.
17 Using binomial with success probability 0.002, the minimum sample size is 1497.
18 796 F. 3d at 66.
19 The data are on page 219 of Vol. 1 of the Joint Appendix.
20 See Table 1 and Table A1.
21 To check the accuracy of the Poisson approximation, the exact probability using the binomial distribution was calculated. It equals 0.002978; in very close agreement with the Poisson approximation.
22 In addition to the small-expected number of American Indian female Finance Officers, their expected number of technical employees was 1.31, clearly less than three. They formed 0.3% of the nations technical workers and the Department employed 377. See the Joint Appendix, Volume I, page 219.
23 Power is the probability for a data-based criterion to detect a significant difference when the difference truly exists.
25 The calculations are based on the binomial distribution, rather than the Poison approximation used previously. There is a large literature concerning various approximations to the sampling distribution of a binomial proportion; see Brown, Cai and Das Gupta (2001) for further discussion.
1. What degree of minority under-representation would be detected if an employer had no minority employees in a workforce of 1500?

2. How large a sample is required in order to have at least 50% chance of a statistically significant result when the odds a minority has are one-half those of a majority applicant?²⁶

When the odds a minority member of the QUALF has of being hired by the employer are the fraction, \( \theta \), of the odds a majority applicant has of being hired, Gastwirth and Greenhouse (1987) showed the employers work force will resemble that of an employer who is hiring from a QUALF with minority availability \( \pi^* \), where \( \pi^* = \frac{\theta \pi}{(1-\pi)+\pi \theta} \). If an American-Indian female applicant had one-half the odds (\( \theta = 0.5 \)) of being hired as a White when \( \pi = .002 \), then \( \pi^* = .001 \). Suppose the State Department had hired 1500 Finance Specialists. Under the alternative model that minorities have one-half the odds of being hired as Whites, one would expect 1.5 of the 1500 to be American-Indian females. The probability of observing no American-Indians among the 1500 is .223. This means that even if American-Indian applicants for the position of Finance Officer had one-half the odds of a White applicant, there is only a 22.3% chance that none of the 1500 would be American-Indian females. This implies that even when the expected number of minority hires is three (under fair practice); there is more than a 75% chance of not detecting a significant shortfall when minority applicants have half the odds of being hired as majority members.²⁷ Thus, one needs a larger sample to detect a meaningful disparity.

In order for the statistical test to have a low probability, say .05, of classifying a fair employer as unfair and a modest power, say 51%, of finding statistical significance if the odds of an American-Indian female being hired are one-half those of a White, the number of officers should be about 4600.²⁸ Then 9.2 minority hires would be expected and a statistically significant shortfall occurs if four or fewer hires were minority. Even in this situation, there is a substantial chance, just under 49%, that one would not detect an employer limiting a minority group to one-half its expected number of positions. In order to have power near 80%, a value often used in the design of medical studies, the expected number of minorities should be 18.2 and a statistically short fall would occur if eleven or fewer minorities are hired.²⁹

For a minority group forming only .2% of the QUALF, the number of employees (or hires) would need to be about 9100, .²⁶Neither the courts nor Congress have established a specific alternative that is important to detect. Because an odds ratio of .5 is considered a medium effect (Olivier and Bell, 2013) it may approximate a manifest imbalance. ²⁷Because statistical evidence of under-representation in disparate treatment cases just places a burden of explaining how the disparity arose from legitimate considerations on the employer, it seems reasonable for courts to check a system in which minorities have only one-half the odds of success as majority members. A firm might consider taking steps to remedy similar levels of minority under-representation in a location or division.

²⁸When \( \pi = 0.002 \) and \( n = 4600 \), \( n \pi = 9.2 \) and the probability of a Poisson random variable with mean 9.2 is less than or equal to four is .0486, very near .05. When \( \pi^* = 0.001 \), \( n = 4600 \), \( n \pi^* = 4.6 \) and the probability of a Poisson random variable with mean 4.6 is less than or equal to four is .513.

²⁹The probability a Poisson random variable with mean 18.2 is less than or equal to 11 is 0.0502 and the probability a Poisson random variable with mean 9.2 is less than or equal to 11 is .783.
an unrealistically large number for most employment settings.\textsuperscript{30} For a minority group forming 5% of the QUALF, when there are 364 hires, the expected number of hires is 18.2.\textsuperscript{31} Then the probability of 11 or fewer minority hires is .0502, when the employers hires are similar to a random sample from the QUALF. On the other hand, if qualified minorities have only one-half the odds of a majority applicant the probability of 11 or fewer hires is .77.\textsuperscript{32} Thus, a statistical test questioning an employer who hires no more than 11 minorities when 18 are expected has a low probability (0.0502) of classifying a fair employer as unfair and a reasonably high probability (0.77) of detecting possible under-representation.

In sum, statistical testing of hiring data when fewer than three minority hires are expected is useless. There is about a 50% chance of detecting noticeable minority under-representation when there are 9 expected hires, but a finding of non-significance is not a strong indicator of fairness. Statistical testing is quite informative when there are more than 18 expected minority hires.

### 3.2. A Composite $T^2$ Test To Test Under-Representation Of Several Minority Groups Simultaneously

Consider testing for under-representation of four protected groups female, Black, Hispanic and Asian. Although females, Blacks, Hispanics and Asians suffered discrimination in the past; the justification for an AAP may not require under-representation in all of them. As long as at least one of these groups is substantially under-represented in the Foreign Service, an AAP is appropriate for the under-represented groups. Let W, B, H, A, I and M denote White, Black, Hispanic, Asian American, American Indian and Male, Female in the subscripts of various variables and quantities. Suppose the State Department is hiring from a large pool of potential employers that may differ from the QUALF, when the proportion of the Caucasian, African American, Hispanic, Asian American and American Indian males are denoted by $P_{WM}, P_{BM}, P_{HM}, P_{AM}, P_{IM}$. Similarly, $P_{WF}, P_{BF}, P_{HF}, P_{AF}, P_{IF}$ denote the corresponding proportions of females. Combining the relevant subgroups yields

\[
P_F = P_{WF} + P_{BF} + P_{HF} + P_{AF} + P_{IF}
\]

\[
P_B = P_{BM} + P_{BF}, \quad P_H = P_{HM} + P_{HF}, \quad P_A = P_{AM} + P_{AF}.
\]

The corresponding minority proportions of the QUALF are denoted by $\pi_{WM}, \pi_{BM}, \pi_{HM}, \pi_{AM}, \pi_{IM}, \pi_{WF}, \pi_{BF}, \pi_{HF}, \pi_{AF}, \pi_{IF}$. Similarly,

\[
\pi_F = \pi_{WF} + \pi_{BF} + \pi_{HF} + \pi_{AF} + \pi_{IF}
\]

\[
\pi_B = \pi_{BM} + \pi_{BF}, \quad \pi_H = \pi_{HM} + \pi_{HF}, \quad \pi_A = \pi_{AM} + \pi_{AF}.
\]

Because the AAP may only be justified when some minority group or females are under-represented, the null hypothesis is

$H_0 : \left( P_F = \pi_F \right) \cap \left( P_B = \pi_B \right) \cap \left( P_H = \pi_H \right) \cap \left( P_A = \pi_A \right)$

and the one-sided alternative hypothesis is:

\[\text{When the odds ratio is } 0.5, \pi^* = 0.02564.\]

\[\text{The number 9100 is obtained by dividing 18.2 by } 0.02.\]

\[\text{The number 364 was chosen as it is in between the number of specialists in level 02 and level 03 in Table A3.}\]
Define \( P \) the difference in the particular category is set to zero. The test statistic is the proportion and the target percentage is used. When this difference is zero or positive, lower than the proportion in the reference population, the difference between the sample being examined for potential under-representation. When the sample proportion is proportions of female, African American, Hispanic and Asian American in the sample responding estimate from the sample by replacing \( P \) with \( P_{ij} \)

\[
H_0: \left( \hat{P}_F - \pi_F \right) \cup \left( \hat{P}_B - \pi_B \right) \cup \left( \hat{P}_H - \pi_H \right) \cup \left( \hat{P}_A - \pi_A \right),
\]

where \( \cap \) and \( \cup \) denote true at the same time and at least one is true, respectively. Define \( \left( \hat{P}_F - \pi_F \right)_- = \min(\hat{P}_F - \pi_F, 0) \), \( \left( \hat{P}_B - \pi_B \right)_- = \min(\hat{P}_B - \pi_B, 0) \), \( \left( \hat{P}_H - \pi_H \right)_- = \min(\hat{P}_H - \pi_H, 0) \), \( \left( \hat{P}_A - \pi_A \right)_- = \min(\hat{P}_A - \pi_A, 0) \), where \( \hat{P}_F, \hat{P}_B, \hat{P}_H, \hat{P}_A \) are the proportions of female, African American, Hispanic and Asian American in the sample detected by simulating its null distribution.

We derive the covariance matrix \( \Sigma \) from the multinomial distribution. The numbers of employees in the ten race-gender categories,

\( (X_{WM}, X_{BM}, X_{HM}, X_{AM}, X_{IM}, X_{WF}, X_{BF}, X_{HF}, X_{AF}, X_{IF}) \), follow a multinomial distribution with probabilities \( (P_{WM}, P_{BM}, P_{HM}, P_{AM}, P_{IM}, P_{WF}, P_{BF}, P_{HF}, P_{AF}, P_{IF}) \) and a 10*10 covariance matrix \( COV \), whose elements are

\[
COV_{ij} = \begin{cases} NP_i(1 - P_i), & \text{if } i = j, \\ -NP_iP_j, & \text{if } i \neq j. \end{cases}
\]

Here \( N \) is the total sample size and \( P_i \) and \( P_j \) are probabilities in the \( i \)th and \( j \)th categories with \( i, j \in (WM, BM, HM, AM, IM, WF, BF, HF, AF, IF) \), respectively. The covariance matrix \( \Sigma = L \ast COV \ast L' \), where

\[
L = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & N^{-1} & N^{-1} & N^{-1} & N^{-1} \\ 0 & N^{-1} & 0 & 0 & 0 & 0 & N^{-1} & 0 & 0 \\ 0 & 0 & N^{-1} & 0 & 0 & 0 & N^{-1} & 0 & 0 \\ 0 & 0 & 0 & N^{-1} & 0 & 0 & 0 & N^{-1} & 0 \end{pmatrix}.
\]

Although Chernoff (1954) and Self and Liang (1987) have developed methods for testing hypotheses where the parameter of interest is a boundary point of the null and alternative hypotheses, they are not directly applicable to the present situation. Therefore, the critical values of the statistic \( T^2 \) were determined by simulating its null distribution.

The number of employees in each race-gender group are provided for 1989 and 1990, most of the employees in 1989 remained in 1990 and the employment patterns were similar in the two years. To avoid testing two highly correlated data sets, we analyze the data from 1990. The employment data for generalists and specialists will be analyzed separately. The AAP in the State department focused on mid-level jobs. Therefore, we combine the two most relevant levels, FS03 and FS02, into one group. Courts may also consider excluding secretaries and clerical jobs from the sample of specialists because their salaries and minimal education requirements are usually lower than other administrative jobs. Thus, analyses of the data, including and excluding the secretary

---

\[\text{The 1989 data were also analyzed and yielded very similar conclusions.}\]
and clerical jobs are carried out to check for under-representation of minorities and females in all specialists jobs and in the higher-level specialist positions.

As we explained in Table 1, the reference race-gender distribution comes from the administrative jobs of the 1980 census because the 1990 census information was not available when the EEOC created the goals for the Department. We also compared the 1990 employee data in the State Department to the demographic mix from the 1990 census.

**Analysis of Gender-Race Composition in State Department**

We obtain the distribution of our modified $T^2$ statistics under null hypothesis empirically by generating multinomial data under the null hypothesis with a fixed sample size of 1000 and calculate $T^2$ from each sample. After generating 15,000 $T^2$ from 15,000 independent samples, one has an empirical distribution of the test statistic under the null hypothesis. Table 2 below listed the empirical quantiles for $T^2$ distributions. Notice that the asymptotic distributions of $T^2$ under the null hypothesis for 1980 and 1990 are slightly different because the mixing coefficients $w_j$, $j = 1, 2, 3, 4$ depend on $(\pi_F, \pi_B, \pi_H, \pi_A)$ which differed in the two Census years (compare the first two rows in Table 1).

Table 2: Empirical Quantiles of under the null (for very large sample sizes)

<table>
<thead>
<tr>
<th>Census</th>
<th>0%</th>
<th>1%</th>
<th>5%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0.36</td>
<td>1.30</td>
<td>2.81</td>
<td>4.69</td>
<td>6.11</td>
<td>9.20</td>
</tr>
<tr>
<td>1990</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.41</td>
<td>1.35</td>
<td>2.84</td>
<td>4.79</td>
<td>6.20</td>
<td>9.29</td>
</tr>
</tbody>
</table>

We apply the one-sided Hotellings $T^2$ test to compare the demographic mix of State Department employees in Foreign Service levels 2 and 3 positions to the national data from 1980 and 1990 Censuses, for generalists and specialists respectively. Furthermore, we test whether the racial composition of generalists in all types of occupations (Administrative, Consular, Economic, Political), as well as the racial composition of specialists excluding those on clerical and secretarial jobs, are the same as that in the national data. The $(\hat{P}_F, \hat{P}_B, \hat{P}_H, \hat{P}_A)$ are the sample proportions of female, Black, Hispanic and Asian in the State Department samples and $(\pi_F, \pi_B, \pi_H, \pi_A)$ are their fractions of QUALF in 1980 and 1990 in Table 1. SAS codes to calculate $T^2$ as well as its asymptotic null distribution given any $(\pi_F, \pi_B, \pi_H, \pi_A)$ and a sample size of N are provided in the Appendix II.

Table 3 reports the degree of under-representation, measured by the difference between fractions of State Department employees and their fraction of the corresponding national data. First, notice that all p-values are less than 0.0001. A p-value less than 0.0001 means that if the race-gender compositions of employees in these types of jobs are considered as a random sample of individuals employed in similar administrative jobs in the nation, the probability of observing such a degree of under-representation in females or minorities is less than 1 in 10,000. Notice that the p-value from the modified
Table 3: The difference between the minority fraction of State Department employees in 1990 and the corresponding fraction of the QUALF in 1990

<table>
<thead>
<tr>
<th></th>
<th>Census</th>
<th>Female</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>$T^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign Service Level 2 and 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generalists 1980</td>
<td>-0.083</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>58.02</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Specialists 1980</td>
<td>-0.178</td>
<td>-0.034</td>
<td>-0.013</td>
<td>0</td>
<td>0</td>
<td>124.21</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Generalists 1990</td>
<td>-0.200</td>
<td>-0.039</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>316.17</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Specialists 1990</td>
<td>-0.295</td>
<td>-0.073</td>
<td>-0.025</td>
<td>0.005</td>
<td></td>
<td>314.55</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>All Occupations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generalists 1980</td>
<td>-0.097</td>
<td>-0.026</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>227.88</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Generalists 1990</td>
<td>-0.214</td>
<td>-0.065</td>
<td>-0.003</td>
<td>0</td>
<td>0</td>
<td>992.02</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Excluding Secretary and Clerical Jobs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specialists 1980</td>
<td>-0.215</td>
<td>-0.032</td>
<td>0</td>
<td>-0.001</td>
<td></td>
<td>1116.18</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Specialists 1990</td>
<td>-0.332</td>
<td>-0.071</td>
<td>-0.012</td>
<td>0.007</td>
<td></td>
<td>2484.91</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

Hotellings $T^2$ test measures the degree of under-representation in at least one minority category, without specifying which categories are significantly under-represented. Whether a specific minority category is under-represented can be determined by examining the differences $(\hat{P}_F - \pi_F)_- , (\hat{P}_B - \pi_B)_- , (\hat{P}_H - \pi_H)_- , (\hat{P}_A - \pi_A)_-$. In the generalists jobs on Foreign Service level 2 and 3, females are under-represented by a difference of 8.3% between their fraction of State Department employees and their fraction of the 1980 national data. (Table 3 Foreign Service level 2 and 3). The other minority groups are not under-represented in those jobs. In the specialists jobs in Foreign Service level 2 and 3 the proportion of Females, African Americans and Hispanics are all less than the corresponding target proportions in the reference population. Notice that there are higher percentages of females and minorities in secretary and clerical jobs that usually have lower ranks (Table A5 in Appendix). Excluding the secretary and clerical jobs from the specialist jobs, females are 21.5% below their target and African Americans and Asians are 3.2% and 0.1% below the 1980 national data, respectively. (Table 3 Excluding secretary and clerical jobs) Clearly, the under-representation of minority groups is more severe after excluding secretarial and clerical jobs. Finally, the minority proportion of individuals employed in administrative jobs in the nation increased noticeably from 1980 to 1990. (Table 1) When the 1990 data were used as the reference, the degree of female and minority under-representation among State Departments employees was greater than when the 1980 data were used. Moreover, under-representation occurred in more job categories, although the degree of underrepresentation of Hispanics in the Generalist occupations appears small.

4. DISCUSSION

At the time the State Department created the AAP, females were severely under-represented in Foreign Service generalists and specialists, especially at the mid-levels
02 and 03. African Americans and Hispanics were also under-represented among specialists. The differences are substantial for the two largest minority groups, females and African Americans. Compared to the 1990 national data, the Hispanic and Asian groups also were usually under-represented. This analysis also supports the courts conclusion that State Departments Affirmative Action was justified. The methodology proposed here can also be used by employers to see whether the demographic mix of their workforce is similar to the QUALF, so they may take appropriate steps to remedy a minority shortfall, e.g. expand their recruitment efforts, before the statistical significance and effect size of the disparity reach a level that would subject them to potential liability.

The use of an effect size measure to supplement conventional hypothesis testing should help employers and courts determine whether a statistically significant disparity is large enough to justify an AAP that does not unduly limit employment opportunities for non-minorities.\textsuperscript{34} Courts have disagreed on whether a meaningful difference, formally measured by an effect size or not, is an appropriate criterion to assess a disparity success or failure rates in disparate treatment or disparate impact class action cases.\textsuperscript{35} If courts or Congress specified the measure of meaningful difference or practical significance that is appropriate for different types of cases, other factors such as the power of the statistical test and the relative costs of the two-types of statistical errors could be used to determine an appropriate level of significance. This might alleviate some of the well-known problems that arise with using .05 as the cut-off for statistical significance in virtually all of science and social science. On the other hand, courts may be reluctant to specify one criterion for practical significance as equal employment cases arise in a wide variety of situations that are not comparable with respect to the number of individuals affected or the risk to the public from an unqualified employee.\textsuperscript{36} Furthermore, courts may be reluctant to create a mathematical criteria for a meaningful difference as employers might act as though employment practices that disadvantage a protected group would not be questioned if the magnitude of the disparity between minority and majority success rates or wages, were less than the specified criteria.

Further study of modifications of conventional testing and alternative approaches,

\textsuperscript{34}In the present case, no White males were laid off or demoted to create a position for a historically under-represented minority. The plaintiffs advancement was delayed but he remained eligible for promotion.

\textsuperscript{35}The opinion in the disparate impact case, \textit{Jones v. Boston} 752 F. 3d 38 (1st Cir. 2014), noted that statistical significance is well-defined and questioned whether it is possible to have a principled definition of “practical” or “meaningful” difference. In contrast the court in \textit{Waisome v. Port Auth. of N.Y. & N.J.}, 948 F.2d 1370, 1376 (2d Cir.1991) found no disparate impact in the passing rates of Black and White applicants where, “though the disparity was found to be statistically significant, it was of limited magnitude”. In that case, however, the difference in the actual promotion rates was not statistically significant. Although the sample sizes in \textit{Waisome} were much smaller than those in \textit{Jones}, the odds ratio measure of “effect size” were similar; .28 for the promotion rates in \textit{Waisome} and .205 for the rates of a positive drug test for African-American applicants to the Police Academy versus Whites. Because of the large sample in \textit{Jones}, the data were statistically significant, even at the .001 level. See Gastwirth (2017) for the discussion of statistical and practical significance in the equal employment cases and references to the related literature.

\textsuperscript{36}See Gastwirth (2017) for cases involving police officers where courts approved employment tests having a disparate impact on a minority group even though the correlation between test score and job performance were low (.21) but statistically significant.
e.g. Bayesian methods, should be helpful to courts in their evaluation of statistical evidence. The data and statistical comparisons in the Shea v. Kerry case are somewhat unusual, which may complicate the development of a suitable prior distribution in the Bayesian context. First, the comparison is between stocks rather than flows, i.e. the employment pattern of the State Department is compared to the minority fractions of the national QUALF. Secondly, because government agencies were only subject to the Civil Rights Act beginning in 1972, a sizeable fraction of the State Departments employees in the late 1980s were pre-Act hires. If one had data comparing the demographic composition of the Departments hires during the period before the AAP was developed, say 1975-1985 to suitably refined Census data, a prior distribution could be based on it. The analysis could be updated every year and the magnitude of the means of the posterior distributions of the hiring odds ratios of the different protected groups at the time the AAP was developed could be examined to see if they reflected substantial under-representation.

Statistical evidence has a limited role in cases concerning a single plaintiff in a disparate treatment case, where the focus is on the treatment (hire or promotion) that the plaintiff received compared with the treatment of similarly qualified majority members. While plaintiffs can support their case with statistics showing a significant difference between the success rates of their group and the majority, unless the disparities are very extreme, the plaintiff needs to submit additional evidence. 37 Similarly, evidence of a balanced workforce at the time the alleged discriminatory event occurred may be helpful to an employer, although it does not immunize an employer from liability for specific acts of discrimination. 38

The EEOC recommended that national data be used by the government agencies to determine the QUALF for various positions at the Department of State. Implicitly, this assumes that all people in the nation at the appropriate skill level have the same probability of applying for a job with the agency; however, a geographically weighted labor market can be determined from applicant data. 39 Because the fraction of residents in the Washington metropolitan area who are African American is higher than their fraction of the nations population, the under-representation of African Americans would likely be more severe if greater weight was given to the local labor force data.

The analysis presented here supports the concerns expressed by Judge Williams when only one minority hire or employee is expected. Indeed, statistical significance cannot be attained unless the expected number of minorities is at least three and

---

37 In McDonnell Douglas Corporation v. Green, 411 U.S. 792, 804-05, the Court said statistics as to petitioner’s employment policy and practice may be helpful to a determination of whether petitioner’s refusal to rehire respondent in this case conformed to a general pattern of discrimination. The plaintiff survived summary judgment in Boone v. Clinton, 675 F.Supp.2d 137, (D.D.C. 2009) where she provided both statistical and anecdotal evidence while the plaintiff in Nicholls v. Philips Semiconductor Mfg., 2011 WL 180565, (S.D.N.Y. 2011) did not because no additional evidence was submitted.


there is only about a fifty percent chance of detecting a meaningful shortfall when nine minorities are expected under fair hiring. When this expected number of minorities reaches 18:

1. A statistically significant shortfall supports the plaintiffs claim, or
2. A non-significant result indicates that the demographic mix of hires or employees is similar to the QUALF, supports the defendants position.

Our findings suggest that when a statistical expert finds a non-significant result, the trial judge should ask about the expected number of minorities and the power of the test. When the expert obtains a statistically significant disparity in a large sample, where the expected number of minorities is also large, a judge might ask the expert whether a significance level less than .05 would be more appropriate.\\footnote{In very large samples, a small difference between the minority proportion of employees or hires and their proportion in the QUALF that may not have legal significance can be statistically significant. See Gastwirth and Xu (2014).}

REFERENCES


https://www.fda.gov/downloads/.../ApprovedProducts/.../UCM457684.pdf


Fleiss, J.L. (1986), Analysis of Data from Multiclinic Trials, Controlled Clinical Trials, 7, 267-275.


APPENDIX I. LEGAL BACKGROUND ON SHEA VS. KERRY

II. The Need for an AAP in the State Department

In 1985, Congress enacted legislation directing the State Department to increase the number of females and minorities in the Foreign Service to be more representative of the national labor force. The State Department implemented a Mid-Level AAP that allowed outside minority applicants, i.e. American Indians, Alaskan natives, Asians, Pacific Islanders, Blacks, and Hispanics, to be hired directly into middle-level positions with higher paygrade. Two years later, Congress was still dissatisfied with the noticeable under-representation of women and minorities in Foreign Service. In 1989, the General Accounting Office released a report from its independent investigation entitled "State Department: Minorities and women are underrepresented in the Foreign Service. The report found (1) the percentages of minorities were still significantly below their corresponding percentages of persons employed in administrative jobs in the nation; (2) although the percentages of minority male had increased, under-representation of females of all race-ethnicity groups was still pervasive. In response, during 1990-1992, the Mid-Level AAP was reinstated by the State Department to address the urgent need of a more representative and diversified officer corps in the United States Foreign. This plan ended around February 1993.

One critical issue concerned the validity of the AAP, i.e., did it satisfy the requirement the US Supreme Court established in Weber and Johnson.41 Ascertaining the existence of a manifest imbalance requires statistical proof of disparities between the racial or gender makeup of the employer’s workforce and that of the qualified labor force. This requires a comparison of the minority percentage of employees in various

41 Supra, n. 1.
jobs with their corresponding percentage of individuals qualified for similar positions in the labor market from which the employer draws its workers.

The statistical underpinnings of the State Department’s AAP consisted of numerical comparisons of the minority and gender percentages of Foreign Service employees in various position with the 1980 Census data. This comparison group was selected by the Equal Employment Opportunity Commission because the Foreign Service recruits from all areas of the nation. At the time the AAP was developed, the 1990 Census data was not available. Because the proportion of the labor force employed in administrative roles held historically disadvantaged groups, especially females, increased during the 1980s, using race-gender data from the 1980 census as reference will under-estimate the minority and female proportions of the qualified labor force available to the Foreign Service in the early 1990s.

Comparing the fractions in the first and last rows of Table 1, it is clear that White males were over-represented and both White and Black females were under-represented in those positions in 1989. Comparing the last row to the 1990 Census confirms these conclusions and suggests that Black males might be under-represented while Hispanic males were not.

The full plan consisted of many similar comparisons between the demographic composition of the professional, administrative, technical, and clerical and a miscellaneous category of other jobs in the Foreign Service with the corresponding 1980 Census data. The data for Foreign Service specialists also indicated a similar pattern of under-representation of White and Black females and over-representation of White males in administrative and technical positions. In clerical positions, White females were over-represented while White males and Blacks, especially females were under-represented.42

The distribution of generalists, specialists and civil service jobs in all ten race-gender categories by ranks in 1989 and 1990 are listed in Tables A1-A6 in the online Appendix III. In the Foreign Service, non-tenured entry-level jobs for generalists are on level FS-07 to FS-04 and non-tenured entry-level jobs for specialists are on levels FS-09 to FS-04. After tenure, an officer may be promoted to mid-levels FS-03 to FS-01. Only after an employee reaches the FS-01 level can they compete for a Senior Foreign Service position. Furthermore, there are three levels in Senior Foreign Service, FE-CM, FE-MC, and FE-OC. In Tables A1-A4, they are denoted by CM, MC and OC.

The Department also submitted data showing that within the senior levels minorities were promoted at a lower rate (3.7%) than White males (8.8%) and White females (9%). They were also promoted from lower to middle level positions at a lower rate than Whites were.43

42The data appear on page 219 of Vol. 1 of the Joint Appendix from USCA case # 13-5153, document 1509794. All references to data from the case will be from the Court of Appeals Record.

43The Joint Appendix in the appellate court record gives the State Departments data and comparisons of its Civil and Foreign Service employees to the national work force. The GAO report presents data indicating that White males had a higher promotion rate to mid-level positions in 1985-1987.
I2. Plaintiff’s Arguments

Mr. William Shea joined the Foreign Service in May 1992 when the Mid-Level AAP was in effect. He filed an administrative grievance with the State Department, claiming that he entered the Foreign Service at a lower pay grade than what would have been the case had he been a minority applicant. Furthermore, he had been receiving less pay with each paycheck than he would be if he had not been discriminated-against as alleged. The State Department denied the grievance. On March 26, 2002, Shea filed a formal complaint of discrimination.

In the senior level jobs (OC, MC, CM), most employees were Caucasian male and the pattern of under-representation of females and minorities is obvious (see Tables A1-A4 in the Appendix). Plaintiff’s analysis focused on Foreign Service level 2 and 3 (FS-03 and FS-02) positions, which were the main focus of the State Department’s AAP and most relevant to Mr. Shea’s position. The plaintiff compared the percentages of the generalist positions at levels 2 and 3 in 1989 held by each racial group to their corresponding percentages of the administrative jobs in the nation. He drew the conclusion that the White percentage was lower than the target, while the percentage of Blacks matched that in the reference population. Hispanics and Asians, on the contrary, are over-represented in the FS-03 and FS-02 jobs at the time.\textsuperscript{44}

- Whites comprised 83.40 percent of FS-03s and FS-02s (1583 of 1898), an underrepresentation compared with their Public Administrator percentage of 86.1.
- Blacks comprised 8.17 percent of FS-03s and FS-02s (155 of 1898), which when rounded to the nearest tenth of a percent, matches their Public Administrator percentage of 8.2.
- Hispanics comprised 5.06 percent of FS-03s and FS-02s (96 of 1898), an overrepresentation compared with their Public Administrator percentage of 3.3.
- Asians comprised 2.69 percent of FS-03s and FS-02s (51 of 1898), again doubling their Public Administrator percentage of 1.3.
- American Indians comprised 0.63 percent of FS-03s and FS-02s (12 of 1898), an underrepresentation compared with their Public Administrator percentage of 0.9.

I3. The Court’s Evaluation Of The Plaintiff’s Statistical Presentation

The District court focused on the under-representation of minorities and women employed as generalists in mid- and senior levels, as Congress had directed the Department to take action to remedy imbalances in those positions. The court agreed with the Department that the data in Table 1 for Foreign Service Generalists showed a manifest imbalance to the disadvantage of White females, Black males and females, Hispanic females and American Indians. Although one might question whether the difference in the representation rates of Hispanic females between the 1989 employment data and

\textsuperscript{44}USCA Case #13-5153 Document #1509794 at pages 134-35 of the Joint Appendix (Vol. 1).
the 1980 Census data is substantial, using 1990 Census data supports the trial judges conclusion.45

The court emphasized that the calculation of racial percentages should not combine males and females and compare the proportions of generalist positions held by members of each race to their corresponding proportion of the relevant national data in judging the appropriateness of an AAP plan. White males were not a group that had suffered historic discrimination. Indeed, Congress had expressed concern about White male over-representation. Since White males were over-represented, while White females were under-represented it is inappropriate to pool the data or apply a combination procedure.46

The court also noted that different comparator groups were employed by the plaintiff and the State Department to justify its AAP.47 Judge Lamberth noted that in Title VII equal employment cases courts require evidence of statistical significance and that the comparator group based the 1980 census would diminish the effect of White under-representation.48 Judge Lamberth further questioned why the plaintiff had still used the 1980 Census data when data that are more recent were available.49 Therefore, the District court found that the State Departments plan satisfied the requirements set down by the Supreme Court in Weber and Johnson. In particular, it did not foreclose advancement opportunities for non-minorities.

The appellate court affirmed the lower courts decision. It emphasized that the statistical evidence of under-representation was quite strong at the senior levels.50 The opinion also referred to non-statistical testimony concerning historical discrimination in recruiting and hiring for the Foreign Service. Finally, it observed that the plan did not overly limit opportunities for non-minorities and was of limited duration.

45See 961 F. Supp. at 37. Although the percentages are small, the Hispanic female proportion of Generalists employed in 1989 (0.008) is slightly less than one-half their proportion in the relevant 1990 national labor force (0.021).

46Methods for determining whether stratified data can be analyzed by combination methods or pooled into a single aggregate data are discussed and illustrated in Miao and Gastwirth (2016) and Gastwirth, Miao and Pan (2017), where references to the statistical literature are provided.


48Ibid. at 47 (noting that Shea continued to use 1980 Census data to determine minority availability even though the GAO report containing that information explicitly questioned their reliability).

49While the opinion refers to a GAO report mentioning this, one can obtain accurate estimates of the demographic composition of many jobs and occupations from the annual average of the monthly Current Population Survey.

50796 F. 3d 42 (D.C. 2015) at 59 (describing the disparity between White and non-White SFS (Senior Foreign Service) employees as overwhelming). The opinion, then recalls the evidence of pervasive historical discrimination in the Foreign Service.
APPENDIX II: SAS CODES TO CALCULATE T2 AND ITS ASYMPTOTIC DISTRIBUTION UNDER NULL

```sas
proc iml;
/*input the percents of (WM, BM, HM, AM, IM, WF, BF, HF, AF, IF) in the
target population, here we use the 1980 census percentages as an example*/
prob = 0.581,0.045,0.022,0.009,0.006,0.282,0.037,0.011,0.004,0.003;
/*input the total sample size, here we use 1000 as an example*/
NumTrials = 1000; /* the number of random realizations of Hotelling's T square
statistics under the null*/
N = 15000;
x = RANDMULTINOMIAL(N,NumTrials,prob);
Cov = -prob#prob; /* replace diagonal elements of Cov with Variance */
Variance = prob#(1-prob);
d = nrow(prob);
do i = 1 to d;
  Cov[i,i] = Variance[i];
end;
testvec = j(N,4,.);
xF = x[,6:10]; pF = xF[,+] / NumTrials; piF = sum(prob[6:10]);
  neg = (pF-piF) # neg;
xB = x[,2:7]; pB = xB[,+] / NumTrials; piB = sum(prob[2:7]);
  neg = (pB-piB) # neg;
xH = x[,3:8]; pH = xH[,+] / NumTrials; piH = sum(prob[3:8]);
  neg = (pH-piH) # neg;
xA = x[,4:9]; pA = xA[,+] / NumTrials; piA = sum(prob[4:9]);
  neg = (pA-piA) # neg;
lineartrans = 0 0 0 0 0 1 1 1 1 1 , 0 1 0 0 0 0 1 0 0 0 ,
  0 0 1 0 0 0 0 1 0 0 , 0 0 0 1 0 0 0 0 1 0 ;
testcov = lineartrans * Cov * lineartrans' / NumTrials;
Tsquare = vecdiag( testvec * inv(testcov) * testvec' );
/*output the 15000 T square values into the dataset nulldist.sas7bdat*/
create nulldist var Tsquare;
append;
close nulldist;
/*calculate the Hotelling's T square test statistic in an observed sample, x is the
numbers of (WM, BM, HM, AM, IM, WF, BF, HF, AF, IF) in the sample*/
x = 607 29 13 8 6 110 9 3 3 0;
NumTrials = sum(x);
testvec = j(4,1,.);
xF = x[,6:10]; pF = xF[,+] / NumTrials; testvec[1] = (pF-piF)*(pF-piF);
```

xB=x[2 7]; pB=xB[+] / NumTrials; testvec[2]=(pB-piB)*(pB-piB);
xH=x[3 8]; pH=xH[+] / NumTrials; testvec[3]=(pH-piH)*(pH-piH);
xA=x[4 9]; pA=xA[+] / NumTrials; testvec[4]=(pA-piA)*(pA-piA);
teststat=testvec'*inv(testcov)*testvec;
print testvec teststat;
quit

Table A1: Generalist, 1989

<table>
<thead>
<tr>
<th>level</th>
<th>White M</th>
<th>White F</th>
<th>Black M</th>
<th>Black F</th>
<th>Hispanic M</th>
<th>Hispanic F</th>
<th>Native M</th>
<th>Native F</th>
<th>Asian M</th>
<th>Asian F</th>
<th>Unsp M</th>
<th>Unsp F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>37</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MC</td>
<td>286</td>
<td>15</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OC</td>
<td>280</td>
<td>22</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>01</td>
<td>605</td>
<td>115</td>
<td>30</td>
<td>10</td>
<td>21</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>02</td>
<td>673</td>
<td>172</td>
<td>54</td>
<td>18</td>
<td>40</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>18</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>03</td>
<td>551</td>
<td>187</td>
<td>45</td>
<td>38</td>
<td>34</td>
<td>13</td>
<td>8</td>
<td>1</td>
<td>16</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>04</td>
<td>718</td>
<td>349</td>
<td>26</td>
<td>26</td>
<td>39</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>28</td>
<td>14</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>05</td>
<td>193</td>
<td>81</td>
<td>4</td>
<td>3</td>
<td>12</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>11</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>06</td>
<td>33</td>
<td>27</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>07</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sum</td>
<td>3377</td>
<td>970</td>
<td>175</td>
<td>98</td>
<td>155</td>
<td>42</td>
<td>16</td>
<td>5</td>
<td>81</td>
<td>38</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Total Generalists 4964

Table A3: Specialist, 1989

<table>
<thead>
<tr>
<th>level</th>
<th>White</th>
<th></th>
<th></th>
<th>Black</th>
<th></th>
<th></th>
<th>Hispanic</th>
<th></th>
<th></th>
<th>Native</th>
<th></th>
<th></th>
<th>Asian</th>
<th></th>
<th></th>
<th>Unsp</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
<td></td>
<td>M</td>
<td>F</td>
<td></td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>MC</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OC</td>
<td>34</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>01</td>
<td>172</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>02</td>
<td>263</td>
<td>35</td>
<td>10</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>03</td>
<td>312</td>
<td>62</td>
<td>22</td>
<td>6</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>04</td>
<td>490</td>
<td>128</td>
<td>20</td>
<td>9</td>
<td>20</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>05</td>
<td>544</td>
<td>239</td>
<td>30</td>
<td>10</td>
<td>23</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>06</td>
<td>231</td>
<td>287</td>
<td>20</td>
<td>19</td>
<td>16</td>
<td>12</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>07</td>
<td>115</td>
<td>271</td>
<td>14</td>
<td>20</td>
<td>7</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>08</td>
<td>44</td>
<td>199</td>
<td>1</td>
<td>15</td>
<td>3</td>
<td>12</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>09</td>
<td>16</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SUM</td>
<td>2236</td>
<td>1240</td>
<td>122</td>
<td>85</td>
<td>85</td>
<td>51</td>
<td>17</td>
<td>7</td>
<td>30</td>
<td>33</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Source: The joint appendix, Volume I, page 212
<table>
<thead>
<tr>
<th>level</th>
<th>White M</th>
<th>White F</th>
<th>Black M</th>
<th>Black F</th>
<th>Hispanic M</th>
<th>Hispanic F</th>
<th>Native M</th>
<th>Native F</th>
<th>Asian M</th>
<th>Asian F</th>
<th>Unsp M</th>
<th>Unsp F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>14</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OC</td>
<td>41</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>01</td>
<td>174</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>02</td>
<td>292</td>
<td>33</td>
<td>10</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>03</td>
<td>315</td>
<td>77</td>
<td>19</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>04</td>
<td>567</td>
<td>133</td>
<td>26</td>
<td>8</td>
<td>24</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>05</td>
<td>448</td>
<td>231</td>
<td>24</td>
<td>11</td>
<td>24</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>06</td>
<td>156</td>
<td>276</td>
<td>16</td>
<td>18</td>
<td>8</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>07</td>
<td>78</td>
<td>258</td>
<td>9</td>
<td>25</td>
<td>4</td>
<td>15</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>08</td>
<td>62</td>
<td>185</td>
<td>0</td>
<td>17</td>
<td>3</td>
<td>13</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>09</td>
<td>23</td>
<td>14</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>SUM</td>
<td>2170</td>
<td>1216</td>
<td>111</td>
<td>91</td>
<td>78</td>
<td>58</td>
<td>19</td>
<td>8</td>
<td>31</td>
<td>35</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

Source: The joint appendix, Volume I, page 213
Table A5: Specialist jobs by finer occupations

<table>
<thead>
<tr>
<th>Occupation</th>
<th>Year</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>Native</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
<td>M</td>
</tr>
<tr>
<td>Professional</td>
<td>89</td>
<td>90</td>
<td>39</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Professional</td>
<td>90</td>
<td>87</td>
<td>40</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Administrative</td>
<td>89</td>
<td>1787</td>
<td>259</td>
<td>92</td>
<td>26</td>
<td>69</td>
</tr>
<tr>
<td>Administrative</td>
<td>90</td>
<td>1706</td>
<td>252</td>
<td>82</td>
<td>27</td>
<td>62</td>
</tr>
<tr>
<td>Technical</td>
<td>89</td>
<td>322</td>
<td>14</td>
<td>19</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>Technical</td>
<td>90</td>
<td>325</td>
<td>12</td>
<td>18</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>Clerical</td>
<td>89</td>
<td>8</td>
<td>911</td>
<td>3</td>
<td>54</td>
<td>0</td>
</tr>
<tr>
<td>Clerical</td>
<td>90</td>
<td>5</td>
<td>886</td>
<td>3</td>
<td>59</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>89</td>
<td>29</td>
<td>17</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>90</td>
<td>47</td>
<td>22</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sec. Engineers</td>
<td>89</td>
<td>124</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Sec. Engineers</td>
<td>90</td>
<td>124</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Sec. Officers</td>
<td>89</td>
<td>677</td>
<td>46</td>
<td>28</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>Sec. Officers</td>
<td>90</td>
<td>662</td>
<td>51</td>
<td>28</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>Secretarial</td>
<td>89</td>
<td>6</td>
<td>908</td>
<td>3</td>
<td>54</td>
<td>0</td>
</tr>
<tr>
<td>Secretarial</td>
<td>90</td>
<td>3</td>
<td>882</td>
<td>3</td>
<td>59</td>
<td>0</td>
</tr>
<tr>
<td>Fin. Officers</td>
<td>89</td>
<td>85</td>
<td>24</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Fin. Officers</td>
<td>90</td>
<td>89</td>
<td>22</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Gen. Services</td>
<td>89</td>
<td>132</td>
<td>20</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Gen. Services</td>
<td>90</td>
<td>131</td>
<td>18</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Comm. Spec.</td>
<td>89</td>
<td>830</td>
<td>122</td>
<td>55</td>
<td>14</td>
<td>45</td>
</tr>
<tr>
<td>Comm. Spec.</td>
<td>90</td>
<td>763</td>
<td>118</td>
<td>46</td>
<td>15</td>
<td>39</td>
</tr>
<tr>
<td>Comm. Tech. Elec</td>
<td>89</td>
<td>196</td>
<td>7</td>
<td>13</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Comm. Tech. Elec</td>
<td>90</td>
<td>199</td>
<td>5</td>
<td>12</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Occupation</th>
<th>Year</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>Native</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
<td>M</td>
</tr>
<tr>
<td>Administrative</td>
<td>89</td>
<td>561</td>
<td>232</td>
<td>49</td>
<td>25</td>
<td>33</td>
</tr>
<tr>
<td>Administrative</td>
<td>90</td>
<td>531</td>
<td>214</td>
<td>49</td>
<td>25</td>
<td>33</td>
</tr>
<tr>
<td>Consular</td>
<td>89</td>
<td>537</td>
<td>270</td>
<td>47</td>
<td>42</td>
<td>25</td>
</tr>
<tr>
<td>Consular</td>
<td>90</td>
<td>529</td>
<td>265</td>
<td>47</td>
<td>39</td>
<td>23</td>
</tr>
<tr>
<td>Economic</td>
<td>89</td>
<td>872</td>
<td>205</td>
<td>19</td>
<td>9</td>
<td>29</td>
</tr>
<tr>
<td>Economic</td>
<td>90</td>
<td>840</td>
<td>203</td>
<td>18</td>
<td>12</td>
<td>32</td>
</tr>
<tr>
<td>Political</td>
<td>89</td>
<td>1346</td>
<td>259</td>
<td>55</td>
<td>21</td>
<td>68</td>
</tr>
<tr>
<td>Political</td>
<td>90</td>
<td>1282</td>
<td>268</td>
<td>52</td>
<td>24</td>
<td>66</td>
</tr>
</tbody>
</table>