- Our mission is to create workplaces free from bias and unlawful discrimination by harnessing the synergies between human resource functions and promoting affirmative action and equal employment regulatory compliance –

Technical Advisory Committee Report on Best Practices in Adverse Impact Analyses

David B. Cohen, M.S., Sr. Vice President

Michael G. Aamodt, Ph.D., Vice President

Eric M. Dunleavy, Ph.D., Senior Consultant

September 2010

THE CENTER FOR CORPORATE EQUALITY (CCE)

1920 I STREET NW, SUITE 400

WASHINGTON, DC 20006

Harold M. Busch

Executive Director

Email: Harold.Busch@cceq.org

Phone: 202-293-2220

www.cceq.org

EXECUTIVE SUMMARY

Although determining whether selection, promotion, and termination decisions result in adverse impact is an important topic for organizations, there is little guidance about the proper way in which these analyses should be conducted. To help provide the EEO community with technical "best practice" guidance on how to conduct adverse impact analyses, the Center for Corporate Equality (CCE) created a Technical Advisory Committee (TAC) consisting of 70 of the nation's top experts in adverse impact analyses and tasked the committee with creating a best practices document. TAC members consisted of a wide variety of EEO experts including industrial-organizational psychologists, labor economists, plaintiff's and defense attorneys, consultants, HR practitioners, and former OFCCP and EEOC officials. The TAC began by creating an extensive survey asking TAC members to indicate how they would handle a variety of data, statistical, and legal issues commonly encountered in conducting adverse impact analyses. The TAC members then gathered at Georgetown University Hotel & Conference Center for a two-day face-to-face meeting to discuss and refine the responses to the survey. At the meeting, TAC members were placed in one of three focus groups: data issues, statistical issues, and legal issues.

Together, expert survey results and discussion from focus groups were used to create this best practice document, which was reviewed in detail by a sub-committee of TAC members. Importantly, this document was distributed without cost to members of the EEO community on September 15, 2010. Some best practice recommendations discussed in the document include the following:

- There is a difference between a job seeker and a job applicant. A job seeker is an individual who expresses an interest in employment with an organization. However, not all job seekers are job applicants for purposes of conducting adverse impact analyses. To be considered an applicant in an adverse impact analysis, TAC members agreed that a job seeker must meet the following five criteria: Express an interest in an open position with an organization, properly follow an organization's rules for applying, meet the basic qualifications for the

position, actually be considered by the organization, and not withdraw from the application process;

- Applicants who submit more than one application for an open position should only be counted once in the adverse impact analysis;
- Organizations should not "guess" the gender or race of applicants who do not self-identify their race and/or gender;
- Job seekers whose credentials were not actually considered by an organization should not be counted as applicants for the purposes of conducting adverse impact analyses;
- Applicants who are offered a job should be counted as a selection regardless of whether they actually accept the offer;
- Measures of both statistical and practical significance should be included in determining the existence of adverse impact;
- There are several measures of statistical significance that can be used to determine adverse impact. The appropriateness of any method is a function of the way in which employment decisions are made;
- The decision to aggregate data across jobs, locations, or requisitions should be made after considering the degree of structure in the selection process as well as the numerical and demographic similarity of the locations and requisitions; and
- Adverse impact analysis of "total minorities" is not a legally appropriate analysis. Instead, the proper analysis is to compare the selection rate of the highest selected group (e.g., Whites, Hispanics) with *each* of the remaining racial/ethnic groups separately.
- There are important differences in pattern or practice scenarios and adverse impact. Disparity analyses may play important roles in both scenarios, but care should be taken to understand the employment processes being analyzed.

TABLE OF CONTENTS

ABOUT THE CENTER FOR CORPORATE EQUALITY

The Center for Corporate Equality (CCE) is a national, non-profit research organization focused on Equal Employment Opportunity. Our mission is to help leaders from various human resource functions identify and use their expertise, understand a breadth of EEO topics, and work together to promote affirmative action and equal employment compliance in their workplaces. Toward this end, CCE conducts research and publishes reports on EEO enforcement, emerging legal topics, and methodological issues.

DISCLAIMER

- This report is not intended to provide specific legal advice. One of the important themes throughout this report is that context matters, and that the lawyer, consultant, or practitioner should consider several factors when planning and conducting adverse impact analyses. For this reason, the information in this report should not be viewed as legal advice. Instead, the information presented in this report should be viewed as a general set of best practices that may or may not appropriately generalize to specific situations. Legal advice depends on the specific facts and circumstances of each individual situation. Those seeking specific legal advice or assistance should contact an attorney as appropriate.

- The results and corresponding recommendations are presented in the aggregate. No TAC member can be explicitly linked to a particular opinion, perspective or recommendation. Survey results were anonymous, and specific focus group participants and comments are confidential. Thus, individual TAC members may disagree with some of the recommendations made in this report, but these recommendations represent the perspective of the majority of the TAC members.

- This report includes best practices at a particular point in time. All data were collected in 2009 and 2010. As such, any changes in EEO statute, scientific literature, and trends in actual practice after 2010 are not captured in this report.

- The information presented in this report is not a criticism of the policies or procedures of any federal agency, specific court rulings, or research from the scholarly literature. Likewise, this is not an attempt to revise the *Uniform Guidelines on Employee Selection Procedures* (UGESP), agency compliance manuals, or any technical authority. Instead, this report includes recommendations for best practices in adverse impact analyses based on data collected from 70 experts in the field.

ACKNOWLEDGEMENTS

This TAC could not have been completed without the hard work and assistance of many individuals. Joanna Colosimo, Johnny Fuller, David Morgan, Fred Satterwhite, and Keli Wilson at DCI Consulting Group Inc. devoted countless hours and energy to various phases of the TAC. Jerilyn Whitmer planned logistics for the on-site TAC meeting and communicated with TAC members on various issues. She also helped review, format and publish this report. She is the main reason why various stages of the TAC ran so smoothly.

In addition, we would like to thank volunteer note takers who documented focus group discussion during the on-site meeting. These volunteers included: Juliet Aiken, Marcelle Clavette, Jeanne Donaghy, Emily Feinberg, Ashley Fulmer, Krystal Larson, Rabiah Muhammad, Monisha Nag, Elizabeth Salmon, Laura Severance, and David Sharrer.

We would like to extend a special note of appreciation to the following organizations that helped sponsor the onsite meeting. These organizations included: Allina Hospitals & Clinics, DCI Consulting Group Inc., ERS Group, Jackson Lewis, Northrop Grumman, and Welch Consulting.

We would also like to thank the Chair of CCE's Board of Directors, David Johnson, for his generous support of this project. We would also like to thank the former and current members of CCE's Board of Directors for their guidance and support during the development of this project, and extend a very special thank you to Harold M. Busch, CCE Executive Director, for his support.

Finally, we would like to thank each and every TAC member for participating. Members volunteered their own time and resources, and in many cases the amount of time and effort was substantial. Some TAC members filled out a 100-item survey, others paid their own way to fly to Washington, D.C. for the on-site meeting, others worked on various sub-committees, and still others reviewed drafts of this detailed report. Some TAC members participated in all phases of the project. CCE realizes that this is a sensitive

topic and in many cases there is little benefit in going “on the record” when there really isn’t a correct answer; we appreciate your willingness and courage to share your expertise. Without you this would not have been possible.

I. Introduction

Although determining whether selection, promotion, and termination decisions result in adverse impact is an important topic for organizations, there is little guidance about the proper way in which these analyses should be conducted. Most books that discuss adverse impact suggest that determining adverse impact is a simple process of comparing the percentage of a focal group (e.g., women) that was selected with the percentage of a comparison group (e.g., men) that was selected and then using either the $4/5^{th}$ rule or a statistical significance test to determine whether adverse impact has occurred.

In the context of this TAC, adverse impact is a general term that describes a meaningful difference in employment decision rates (e.g., hires, promotions, terminations) across two groups (e.g., men and women). A "disparity" often stems from a specific and facially neutral selection procedure (e.g., an employment test, an interview, a resume screen). In situations where there are no identifiable steps in an unstructured selection process, similar analyses of disparity are used to identify differences as indirect evidence of discrimination, and are often combined with anecdotal evidence of intentional discrimination. This is referred to as a disparate treatment pattern or practice scenario. Adverse impact is often presented as a difference in employment decision rates across two groups. For example, 50 of 100 male applicants may have been hired, while 30 of 100 female applicants may have been hired. Is the male hiring rate (50%) meaningfully different than the female hiring rate (30%)?

This report focuses on disparity stemming from some form of applicant flow data. In circumstances where applicant flow data are not available, an adverse inference theory may be applied, and workforce composition or hiring rates can be compared to some form of census data representing availability. Although many of the topics discussed in this report are applicable to these availability analyses, CCE intentionally chose to focus on situations where actual applicant flow data are available, and the probative question of interest is whether there is a meaningful difference in actual employment decision rates across two groups.

When adverse impact exists, the organization using a selection procedure of interest must justify its use, often via evidence of job relatedness using validity research. Likewise, in a traditional impact case, the organization is also generally required to consider alternative selection procedures (i.e., tests tapping different constructs) or alternative testing methods (e.g., video-based vs. paper-pencil testing) that would have been equally valid yet less adverse on members of the impacted group. If the organization cannot justify the use of a selection procedure or show that it considered alternatives, the use of that selection procedure may be deemed discriminatory.

Analysts who conduct bottom line (i.e., applicant to hire) adverse impact analyses on a regular basis are keenly aware of the many difficult decisions that go into determining which job seekers are actually going to be included in the analysis and which statistical tests should be used to determine adverse impact. These difficult decisions often result in analyses conducted by one party (e.g., an employer) that provide a different conclusion than that of another party (e.g., EEOC, OFCCP, private litigants). Often, these differing conclusions are the result of legitimate differences in opinion regarding how adverse impact analyses should be conducted. There are other times, however, when the decisions made by each party may be more strategic than philosophical.

As a result of the differing opinions that exist regarding the "proper" way to conduct adverse impact analyses, in December of 2009, the Center for Corporate Equality (CCE) assembled a Technical Advisory Committee (TAC) that included 70 of the nation's top experts on adverse impact. The purpose of the TAC was to create a present day "best practices" document on how to conduct adverse impact analyses.

CCE took a four-phase approach to the TAC process. First, CCE invited a group of experts who are involved with the equal employment opportunity community to participate. The initial list of potential members represented a wide range of backgrounds, perspectives, and areas of expertise. This list included lawyers, labor economists, statisticians, industrial/organizational psychologists, ex-government officials, and EEO practitioners. A careful mix of experts representing both plaintiff and management were

among those invited. An overwhelming majority of invitees accepted the invitation. Next, CCE developed a comprehensive survey to establish some initial expert opinion on basic concepts, and to identify issues where strong differences and agreement in opinion exist. This survey was vetted by a sub-committee of TAC members. The survey included sections regarding respondent backgrounds, gathering and preparing data, statistical methods for data analysis, and legal/policy issues. The results of this survey served as a bridge to phase three of the process, where a two day in-person focus group session was held at Georgetown University Hotel & Conference Center in Washington, D.C. These focus groups provided the opportunity for members to discuss the survey results and issues of disagreement in detail. This report represents the fourth phase of the process, where a written report summarizes results. The report was vetted by a sub-committee of TAC members to ensure that previous phases of the TAC process are accurately presented.

It is important to note that the results and recommendations in this report represent the opinions of a majority of TAC members on a particular issue. Importantly, no single TAC member can be singled out as endorsing a particular perspective. In fact, it is almost certain that no TAC member will personally agree with 100% of the recommendations found in this report. Some TAC members contested certain survey items, and other TAC members disagreed with a majority of members that participated in focus groups. Thus, the recommendations reported here are suggested best practices generally agreed upon by a majority of TAC members, and should be reported and interpreted as such.

The Technical Advisory Committee

As shown in Table 1.1, the 70 TAC members represented a wide variety of professions and perspectives including I/O psychologists, labor economists, defense attorneys, plaintiff's attorneys, HR practitioners, academicians, statisticians, and former OFCCP and EEOC directors. A majority of the TAC members have worked as an expert on both the plaintiff and management sides of an adverse impact case. TAC members had an average of 21 years experience in conducting adverse impact analyses. Most TAC

members had experience working with organizations in both the public and private sectors. A complete list of the TAC members can be found in Table 1.2.

Table 1.1: TAC Member Characteristics

Characteristic	Percent
Primary Area of Expertise	
I/O psychology	38.0
Employment law	27.0
Labor economics	15.9
HR compliance	14.3
HR statistics	4.8
Current Employment Position	
External consultant	47.1
Attorney	24.3
Internal practitioner	15.7
Academic	12.9
Gender	
Male	68.6
Female	31.4

Table 1.2: Technical Advisory Committee (TAC) Members

Michael Aamodt, DCI Consulting Group, Inc.
Nancy Abell, Paul Hastings
Nikki Alphonse, Northrop Grumman
Pete Aponte, Abbott
Mary Baker, ERS Group
Nita Beecher, ORC Worldwide
Vaida Bell, IBM
Elizabeth Owens Bille, Society for Human Resource Management (SHRM)
Jim Bluemond, Northrop Grumman
Harry Brull, Personnel Decisions International
Inderdeep Chatrath, Duke University
David Cohen, (Chair), DCI Consulting Group, Inc.
Deb Cohen, Society for Human Resource Management (SHRM)
Virginia Connors, FedEx
David Copus, Ogletree Deakins
Donald Deere, Welch Consulting
Tanya Delany, IBM
Aaron Dettling, Balch & Bingham, L.L.P.
Cari Dominguez, Dominguez & Associates
Bill Doyle, Morgan Lewis
Dennis Doverspike, University of Akron
Eric Dunleavy, DCI Consulting Group, Inc.

Burt Fishman, Fortney & Scott, LLC
Christopher Erath, NERA Economic Consulting
David Fortney, Fortney & Scott, LLC
John Fox, Fox, Wang, & Morgan
Jon Geier, Paul Hastings
Barry Goldstein, Goldstein, Demchak, Baller, Borgen & Dardarian
Dana Glenn-Dunleavy, Association of American Medical Colleges
Irv Goldstein, University of Maryland
Art Gutman, Florida Institute of Technology
Paul Hanges, University of Maryland
Joan Haworth, ERS Group
Lisa Harpe, Peopleclick
Valerie Hoffman, Seyfarth Shaw
Alissa Horvitz, Littler Mendelson
David Johnson, Allina Hospitals and Clinics
Joe Kennedy, JN Kennedy Group, LLP
Joe Lakis, Norris, Tysse, Lampley & Lakis, LLP
David Lamoreaux, Charles River Associates
Tanya Lewis, Raytheon
Mort McPhail, Valtera Corporation
Lilly Lin, Developmental Dimensions International (DDI)
Cyrus Mehri, Mehri & Skalet
Scott Morris, Illinois Institute of Technology
Fred Melkey, The Hartford
Lorin Mueller, American Institutes for Research
Patrick Nooren, Biddle Consulting Group, Inc.
Kevin Murphy, Penn State University
Dwayne Norris, American Institutes for Research
Chris Northup, Ellen Shong & Associates, LLC
Scott Oppler, Association of American Medical Colleges
James Outtz, Outtz and Associates
Ramona L. Paetzold, Texas A & M University
Doug Reynolds, Developmental Dimensions International (DDI)
Paul Sackett, University of Minnesota
Lance Seberhagen, Seberhagen & Associates
Ellen Shong-Bergman, Ellen Shong & Associates, LLC
Mickey Silberman, Jackson Lewis LLP
Murray Simpson, Peopleclick
Evan Sinar, Developmental Dimensions International (DDI)
Joanne Snow, JSA Consulting, Inc.
Andy Solomonson, Previsor
Matthew Thompson, Charles River Associates
Michael Ward, Welch Consulting

Determining Best Practices

TAC Member Survey

The first step in developing the best practices document was to develop a detailed survey to be completed by TAC members. The initial survey was created by Eric Dunleavy, David Cohen, David Morgan, and Michael Aamodt and then reviewed by a subcommittee. The survey asked TAC members how they would handle both common and unusual situations that occur when conducting adverse impact analyses. The survey was sent to TAC members on October 21, 2009 and 64 of the 70 TAC members returned the survey. A complete copy of the survey questions and results can be found in Appendix A.

The survey was organized into three content domains: gathering and preparing data, statistical methods for analysis, and legal/policy issues. Note that some TAC members have very specialized EEO expertise, and were instructed not to answer survey questions if the topic was outside of their expertise. At the beginning of a new survey section, members were asked if they felt that they had expertise in that particular content domain. Generally those who did not feel that they had expertise in a particular area skipped that section of the survey.

In-Person Focus Groups

After the survey results were tabulated, 45 TAC members met in person for two days at Georgetown University Hotel & Conference Center in Washington, D.C. TAC members were placed into one of three focus groups to discuss the survey results: gathering and

preparing data, statistical methods for data analysis, and legal/policy issues. During these focus groups, TAC members discussed the survey results to identify and understand any points of disagreement. For several of the survey questions, the focus groups discovered that agreement rates would have been higher had a question been worded differently or had the person completing the survey not read more into a question than was being asked. In addition, other issues that were not specifically asked on the survey were discussed during the focus group meetings.

To ensure that the quality of the focus group discussions was properly captured, multiple note-takers were present to record the key points being made as well as the ultimate best practice, if any, suggested by the focus groups. The note-takers were CCE staff and I/O psychology graduate students from the University of Maryland and Radford University.

Best Practice Document

Following the in-person focus groups, the survey results and focus-group results were summarized and organized into this best practices document. The document was then reviewed by another subcommittee to ensure that it properly captured the general agreement of the TAC.

The following pages contain what TAC members generally agree are suggested best practices for conducting adverse impact analyses. Every attempt was made to appropriately weight survey results and focus group discussion in making recommendations. When reading the document, you will often see a percentage in parentheses following a statement. This number represents the percentage of TAC members whose survey responses indicated they agreed with the statement. It is safe to say that, following the clarification and discussion that occurred in the focus groups, these percentages probably understate the actual level of expert agreement. With that said, it is important to note that even on issues in which there was a high level of agreement, there was often a strongly stated minority opinion that disagreed with the majority opinion.

II: Best Practices: Gathering and Preparing Data for the Analysis

The basic idea in conducting a bottom line (i.e., applicant to hire) adverse impact analysis of applicant flow data for an overall selection process is to compare the percentage of applicants selected from one group (e.g., men) with the percentage of applicants selected from another group (e.g., women). Regardless of the type of statistical analysis ultimately used to compare these percentages (e.g., 4/5th rule, standard deviation test, Fisher's exact test), the first step in the analysis process is to determine the number of applicants and the number of individuals hired/selected for each group. Although it would appear that this first step should be a simple one, determining who is an applicant and who is selected can be a contentious process.

In this section, we will:

- Explain the criteria for a job seeker to be considered an "applicant";
- Discuss how to handle multiple applications submitted by the same job seeker;
- Review how to handle special applicant populations (e.g., contract employees);
- Discuss how to handle job seekers who submitted applications but were never actually considered by the organization;
- Explain how to handle applicants who withdrew from the process;
- Discuss the criteria for an applicant to be considered "selected"; and
- Consider two additional areas of concern: applicants who do not self-identify race or gender and requisitions that remain open at the end of an Affirmative Action Plan (AAP) year.

Criteria for a Job Seeker to be Considered a Job Applicant

A job seeker is an individual who expresses an interest in employment with an organization. However, not all job seekers are job applicants for purposes of conducting

adverse impact analyses.[1] To be considered an applicant in an adverse impact analysis, TAC members agreed that a job seeker must meet the following three criteria:

Express an interest in an open position with an organization

To be considered as an applicant, a job seeker must apply for a specific position with an organization (96%).[2] Job seekers who apply after an external candidate has been hired (100%) or whose application had not been reviewed prior to an internal candidate being selected for the position (79%) are not considered applicants and are excluded from the bottom-line adverse impact analysis.

Properly follow an organization's rules for applying

A job seeker must submit an application/resume in the manner requested by the organization and must follow any organizational rules for applying. That is, if the organization, as a general rule, requires job seekers to submit a resume on-line and the job seeker mails his/her resume, the job seeker would not be considered an applicant and thus would be excluded from the adverse impact analysis.[3]

Examples of job seeker behaviors that might merit exclusion include:

- Not applying for a specific position (96%). If, however, an organization considers a job seeker for a position, even though the job seeker did not apply for that specific position, the job seeker should be included in the analysis (94%);
- Not answering an application question related to determining whether the applicant meets the basic qualifications for the job (83%);

[1] It is important to note that EEOC defines an applicant in accordance with the UGESP Questions and Answers document (Equal Employment Opportunity Commission et al., 1979). OFCCP on the other hand, relies on their Internet Applicant Regulations (Internet Applicant Rule, 2005). The definition discussed here is that of the TAC.

[2] Recall that this number in parentheses represents the percentage of TAC members whose survey responses indicated they agreed with the statement.

[3] Although employers are expected to permit individuals with disabilities who require accommodation in the application process to mail in a resume, we are not referring to an accommodation exception.

- Not following the formal application process (75%);
- Submitting an application that could not be read either because of poor penmanship or the application was not completed in English (71%); and
- Not signing an application as required (65%).

Recommendation 2.1: *As a best practice, organizations:*

- *Have the right to determine what constitutes a complete application;*
- *Must be consistent in applying rules that exclude applicants;*
- *Should communicate which sections of an application must be completed as well as the consequences for not completing the required sections. This is especially important for organizations that use one application form for all jobs, as not every question on the application will be essential for a given job;*
- *Should exclude applicants for not completing a section only if that section is important for determining job related information (e.g., experience) or if completion of that section has legal implications (e.g., a signature that attests to the accuracy of the application information);*
- *Should train recruiters and other decision makers regarding the importance of consistency in using screening procedures; and*
- *Should monitor decision makers to ensure that screening procedures are being properly followed.*

Meet the basic qualifications for the position

Job seekers should not be considered applicants and will be excluded from the bottom-line adverse impact analysis if they do not meet the basic qualifications required for the position. Examples of such situations include:

- Not being eligible to work in the United States because they lack the required work visa (87%);
- Not meeting the basic qualifications (e.g., degrees, certifications, experience) for hire (84%);

- Being former employees who do not meet the company requirements to be rehired (e.g., terminated, applied too soon after being terminated; 80%);
- Being current employees who have not met minimum tenure in the current job or have documented performance problems; and
- Providing false information about their basic qualifications (72%).

For a selection criterion to be considered a basic qualification, it must be formally stated by the organization and consistently applied to job seekers. The term "consistently applied" does not require that the criterion be applied 100% of the time, but exceptions to the criterion must be rare and the result of extenuating circumstances (e.g., need to fill the position although no job seekers completely met all basic qualifications; unusual life experience that compensates for lack of basic qualification). Basic qualifications cannot be preferences; they must be a standard that if not met, indicates the job seeker could not perform the job. Likewise, basic qualifications must be non-comparative, such that applicants either have the qualification or do not, and no judgment is made about one applicant having more of a qualification than another. As a result, two individuals with knowledge of the job reviewing an applicant's qualifications should come to the same conclusion.

Recommendation 2.2: *Only job seekers meeting the basic qualifications for a position should be considered as applicants in adverse impact analyses. As such, it is essential that the organization be consistent in its application of these basic qualifications.*

Multiple Applications from the Same Job Seeker

It is not unusual for the same job seeker to submit multiple applications for the same job requisition or for an open and continuous requisition. In both cases, TAC members agreed that the applicant should be counted only once (93%). If the applicant received a job offer, the application resulting in the job offer should be used. If the applicant did not receive a job offer, the most recent application should be used. The remaining applications should be kept on file but should not be included in the adverse impact calculations.

A more complicated situation occurs when an applicant submits a separate application for several separate requisitions. If the adverse impact analyses are conducted separately for each requisition, each application should be included in the analysis (one application for each analysis). TAC members differed in how the above situation should be handled when the requisitions are aggregated into one analysis. A majority (58%) of TAC members responding to the survey thought that each application should be counted, but focus group members were more split on how such applications should be counted.

Recommendation 2.3: *Applicants who submit more than one application for the same requisition should only be counted once in the adverse impact analyses. Copies of all applications, however, should be retained for record keeping purposes.*

Special Applicant Populations

For certain types of job openings, applicants should not be considered in an adverse impact analysis. Such openings include:

- Temporary workers, not on the employer's payroll, hired from an outside firm (86.0%);
- Contract employees such as those under a 1099 provision (80.4%);
- Employees working outside the United States (77.8%); and
- Interns (65.2%).

TAC members were not able to reach general agreement regarding temporary workers who are on the employer's payroll, as 52.2% of TAC members thought they should be excluded from the analysis and 47.8% thought they should be included.

Consideration Issues

There are often situations in which an employer receives a large number of job seekers for a relatively small number of positions. In such situations, employers often use "data management techniques" in which they will review only a sample of the applications. For example, an employer might receive 1,000 applications for an opening and randomly or sequentially select 50 to review. If qualified applicants are not found in the first 50, another 50 applications will be randomly selected to review.

In such situations, the question arises regarding whether the employer must include all 1,000 applications in the adverse impact analyses or just those that were considered. The general agreement of the expert focus group was that, as long as the data management technique is random and used systematically, only those applications that were *considered* should be included in the analysis. Therefore, in the example in the preceding paragraph, only the 100 applicants (two rounds of 50) actually reviewed would be included in the bottom line adverse impact analysis. If, however, the data management technique uses an algorithm to evaluate or rank the applications, all job seekers that meet the three criteria discussed in the previous section should be considered applicants and included in the adverse impact analysis. This observation has particular implication for electronic "Job Board" searches, which rank and prioritize responses based on "relevancy" algorithms proprietary to each company supporting its board (i.e., HotJobs may use a different sorting and priority algorithm than will Monster.com).

A related issue occurs when a resume arrives after the initial screening period, but was never opened or considered. The general agreement of the expert focus group was that such job seekers should not be included as applicants in the adverse impact analysis.

Recommendation 2.4: *Job seekers whose application materials were not considered by an organization should not be counted as applicants during an adverse impact analysis. This recommendation does not apply to situations in which the organization used a data management technique to evaluate or rank a job seeker's qualifications.*

Applicant Withdrawals

Job seekers who withdraw from the application process are not counted as applicants in bottom line adverse impact analyses, but may be considered an applicant for a particular step in the employer's selection process. Applicant withdrawals typically come in two forms: formal withdrawal or implied withdrawal. Formal withdrawal behaviors include applicants who directly communicate to the organization that they are no longer interested in the position (89%). Implied withdrawal behaviors include:

- Not providing a valid means of contacting the applicant (89%). Examples would include not putting a phone number on the application or providing an incorrect email address;
- Not returning company emails or calls (91%);
- Not showing up for a scheduled interview or testing session (87%);
- Being unwilling to perform job-related travel (85%) or being unwilling to work a required shift or start date (81%); and
- Expressing salary requirements that are too high for the position in question (73%).

Recommendation 2.5: *Job seekers who formally withdraw from the selection process, do not provide valid contact information, or whose actions suggest they are no longer interested in the position should be considered to have withdrawn from the process and should not be considered as applicants for the purposes of conducting adverse impact analyses.*

Criteria for Being Considered a Selection

In the previous few pages, we discussed who should be considered an applicant (i.e., included in the denominator in adverse impact calculations). An equally important issue is who should be considered a *selection* (i.e., included in the numerator in adverse impact calculations). Although it is obvious that a person who accepts a job offer and reports for

work is a selection (100%), there are other situations in which an applicant should be considered as a selection to include in the numerator of the adverse impact analyses. These situations include applicants that:

- Accept a job offer but do not report to work (78%);
- Are offered a job but decline the offer (77%); and
- Accept a conditional job offer but do not report for a post-offer medical exam (69%).

TAC members who did not think the above three situations should be considered selections thought that the applicants should be treated as having withdrawn from the process and should not even be counted as applicants.

TAC members had difficulty in reaching agreement regarding how to handle conditional offers of hire. As shown in Table 2.1, a plurality of members thought that applicants failing a post-offer drug screen should be considered as a "selection" whereas a plurality thought that applicants failing post-offer medical exams, physical ability tests, and background checks should be considered as rejections. Only 7% of TAC members thought that applicants failing post-offer tests should be considered as having withdrawn from the process. During the expert focus group sessions, general agreement was reached that if the post-offer test is required by law (e.g., drug screen) the applicant should be counted as a selection in bottom line analyses.

TAC members were in agreement that applicants failing background checks (83%), physical ability tests (83%), and drug tests (80%) that were made prior to a conditional offer of hire are to be considered rejections.

Table 2.1: Failure of Post-Offer Exams

Issue	Applicant Disposition			
	Selected	Rejected	Withdrawn	Response Count
Conditionally offered job but failed drug test	50.00%	42.60%	7.40%	54
Conditionally offered job but failed medical exam	42.60%	50.00%	7.40%	54
Conditionally offered job but failed physical ability test	37.00%	55.60%	7.40%	54
Conditionally offered job but failed background/credit check	35.20%	57.40%	7.40%	54

Recommendation 2.6*:* *Applicants who are offered a job should be counted as a selection regardless of whether they actually accept the offer as should applicants who accept a conditional job offer but do not report for a post-offer medical exam.*

Other Issues of Concern

Applicant Identification Issues

As part of the application process, job seekers are asked to self-identify their gender and race/ethnicity. Although the vast majority of job seekers provide this information, some do not. In such cases, the question becomes, how do you handle applicants in the analysis when their gender or race is unknown? For applicants who are not hired, there is no alternative source for identification. For applicants who are hired, the organization can obtain their gender or race after hire; either through self-identification (the preferred method endorsed by the EEOC and OFCCP) or through visual identification (a last resort). It is important to note that obtaining race/ethnicity and gender information for employees goes beyond conducting adverse impact analyses as organizations that are required to submit an annual EEO-1 report are required to report on the race/ethnicity and gender for every employee[4].

[4] Note that colleges and universities are required to fill out and submit an annual Integrated Postsecondary Education Data Systems (IPEDS) report. Unlike the EEO-1 report, the IPEDS report allows for unknowns in the employment data file, and for this reason backfilling race/ethnicity may not be possible.

There is some disagreement regarding whether an organization should take the gender and race information of new hires and "backfill" the applicant records missing such information. The advantage to such a practice is more complete information. The disadvantage is that by only backfilling the information from applicants who were hired, a distorted picture of the applicant pool might occur. As shown in Table 2.2, the majority of TAC members (82%) completing the survey thought organizations should backfill from the hires file. Such a conclusion was also reached by the focus group discussing data issues.

Table 2.2: How do you handle applicants who do not self-identify during the application process for gender and/or race/ethnicity but subsequently provide that information after they are hired?

Response	% Agreeing	Response Count
Backfill from hires file into all requisitions to which they applied	54.00%	27
Backfill from the hires file only into the requisition to which they were hired, and keep as missing in any other requisitions to which they may have applied	28.00%	14
Other	10.00%	5
Keep as missing in all requisitions to which they have applied	8.00%	4

There was, however, a very vocal minority opinion that backfilling from the hires file should not occur in any statistical analysis of applicant flow. This group suggested that backfilling from hire records will bias the statistical calculation of hiring rates for the group that is more likely not to self-identify. This may inflate a disparity when none

exists or may obscure a real disparity.[5]

Agreement was not so clear regarding applicants who were not hired but for whom the organization has race/ethnicity information from other sources (e.g., other applications). As shown in Table 2.3, the majority (57%) of TAC members thought that backfilling should occur for these applicants as well.

Table 2.3: How do you handle applicants who were not hired but for whom you have race/ethnicity information from other sources?

Response	% Agreeing	Response Count
Backfill from other sources into all requisitions to which they applied	56.90%	29
Keep as missing in all requisitions to which they have applied	25.50%	13
Other	17.60%	9

TAC members were clear that an organization should not provide a "best guess" on missing race/ethnicity (96.1%) or gender (82.0%) information based on an applicant's name. A narrow majority of TAC members (52%) thought that it was a reasonable practice to visually identify an applicant's gender/race/ethnicity information during an in-person interview. In the focus group, TAC members agreed that, because self-identification is a voluntary process for the applicant, an organization is under no legal

[5] The reasoning of this group is as follows. Consider the following two examples. Suppose that an applicant pool contains 50 Whites and 50 Blacks, but that 25 of the Whites do not identify their race. Suppose that eight Whites (four who identified their race and four who did not) and eight Blacks are hired. Thus hiring is strictly proportional with regard to race and this is evident from the statistical comparison of the 4/25 hiring rate for identified Whites and the 8/50 hiring rate for Blacks. If, however, race for the four unidentified but hired Whites is backfilled, then the statistics mistakenly suggest over-hiring of Whites: 8/29 vs. 8/50.
As a second example, suppose an applicant pool contains 50 Whites and 50 Blacks, but that 25 of the Blacks do not identify their race. Suppose that 10 Whites and six Blacks (three who identified their race and three who did not) are hired. Now, hiring actually favors Whites as shown by the statistical comparison of the 10/50 for Whites and the 3/25 for identified Blacks. If, however, race for the three unidentified but hired Blacks is backfilled, then the statistics mistakenly suggest a statistically non-significant over-hiring of Blacks: 10/50 vs. 6/28.

obligation to visually identify the applicants who chose not to self-identify. Focus group members acknowledged that, should visual identification be used, care should be taken to not only ensure accuracy, but that the visual identification process does not result in "too much attention" being placed on an applicant's gender or race.

Several focus group members questioned whether allowing recruiters to visually identify applicants might result in biased data. This is especially a concern in situations in which recruiters are rewarded for diverse applicant pools.

Recommendation 2.7*: Organizations should not "guess" the race or gender of applicants who do not self-identify. Although organizations are not legally required to backfill the race or gender of applicants who are hired, doing so is a reasonable practice.*

Open Requisitions at the End of the AAP Year

When completing an annual affirmative action plan (AAP), it is common for an employer to complete the AAP year with requisitions for which no hire has yet been made. In such situations, the majority (69.2%) of TAC members thought the applicant flow data should be counted in the year in which an employment decision was made or in which the requisition was closed. Some TAC members thought that some of the activity could be included in a steps analysis in the prior year but that the bottom line analysis should be included in the year in which the hiring decision was made. Only a small percentage (7.7%) of TAC members thought the applicant flow data should be counted in the year in which the applications were received.

Every member of the focus group agreed that the applicant flow data should be counted in the year in which the hire is made. Focus group members generally agreed that the "hire" occurs on the day in which the hiring decision is made, although other dates (e.g., date offered, date offer was accepted; date employee is added to payroll) are also

acceptable. Focus group members advised that, regardless of the data used, organizations should be consistent in how they define the "hire" date.

Recommendation 2.8*: Applicant flow data should be counted in the year in which the hire is made rather than the year in which the application was received.*

Bottom Line Analyses Versus Step/Component Analyses

When determining adverse impact, there are often two separate analyses that are conducted: a bottom-line analysis (i.e., the number of selections divided by the total number of applicants) followed by a series of step/component analyses (i.e., the number of job seekers passing a step divided by the number of job seekers participating in that step). The bottom-line analysis determines whether the selection, promotion, or termination process *as a whole* results in adverse impact. A step analysis focuses on the adverse impact from a particular component in the selection process. Typically a step analysis is necessary when there is substantial adverse impact in the total selection process and that process can be broken down into individual components (UGESP, 1978), or when a claim is made that a specific step in the process is discriminatory (e.g., *Connecticut vs. Teal*, 1982).

For example, consider a situation in which an employer's selection process consists of three steps/components: an application screen, an interview, and a physical ability test. The bottom-line analysis indicates that the selection process has adverse impact against women. The next step is to run adverse impact analyses on each of the three steps/components to determine which of the three is causing the adverse impact.

From a data perspective, the number of applicants and hires in a bottom-line analysis will be different from the number of applicants and "hires" in a step analysis. That is, using the previous example, an applicant who does not show up for an interview would be considered a withdrawal for the bottom-line analysis but would be considered an applicant and a "pass" for the step analysis conducted on the application screen

component. In the above example, the applicant would also be considered as a withdrawal for the interview step analysis as well as for the physical ability test step-analysis.

It was the general agreement of TAC members that, if a job seeker does not meet the definition of an applicant described in the previous section, the job seeker is not included in the bottom-line analysis and may also be removed from some step/component analyses. If an applicant withdrew during any stage in the process, he/she is excluded from the bottom-line analysis and any steps/components in which he/she did not participate but is included in the analyses of the steps/components in which he/she participated.

It is important to note that if a step or component results in adverse impact, the employer must demonstrate the validity of the step or component, regardless of whether there is adverse impact in the bottom-line analysis (*Connecticut vs. Teal*, 1982).

Summary

This section of the TAC obviously covered many important issues, and all of these issues may drastically affect how an adverse impact analysis is structured. For example, whether data mirror the reality of the actual employment decision-making process under scrutiny, and what conclusions can be made based upon the results of the analysis. At the end of the focus group participants were asked what themes they thought were most important in this section. The following four themes emerged:

1. It is essential to create requisition codes for each job opening (this may include multiple hires within the requisition) and to indicate the disposition of each job seeker;

2. Not all job seekers will be considered applicants for the purpose of conducting adverse impact analyses. To be considered an applicant, a job seeker must express an interest in

an open position, meet the minimum qualifications for the position, and follow the organization's rules for applying;

3. Job seekers who withdraw formally (e.g., inform the organization that they are no longer interested in the position) or informally (e.g., do not return calls, fail to show for an interview) should not be considered as applicants in the adverse impact analyses; and

4. Applicants who are offered a position but either decline the position or do not report to work should be considered as "selections" in adverse impact analyses.

III: Best Practices: Statistical Methods for Adverse Impact Analyses

Once applicant flow data have been cleaned and refined to determine the applicant pools, the next step is to conduct some form of statistical analysis. This analysis usually attempts to differentiate a meaningful disparity from a trivial disparity, and may utilize a number of paradigms and consider any number of contextual factors. This chapter summarizes TAC member opinions on a number of statistical issues, and both survey results and focus group discussions are highlighted as part of each recommendation.

Although statistical analyses of disparity have received recent attention in scholarly literature (e.g., Collins & Morris, 2008; Bobko, Roth, & Switzer, 2006), book chapters (e.g., Zedeck, 2010; Bobko & Roth, 2010; Siskin & Trippi, 2005) and equal employment opportunity texts (e.g., Biddle, 2005; Gutman, Koppes, & Vodanovich, 2010), no resource provides a comprehensive and present day treatment of the myriad practical issues considered in this chapter. In this section, we will consider:

- Different paradigms for conducting analyses of disparity;
- Best practices for using statistical significance tests to assess disparity;
- Best practices for using practical significance tests to assess disparity;
- Appropriate methods to aggregate/disaggregate data; and
- More complex analyses of disparity.

Paradigms for Disparity Analyses

There are a number of different paradigms available for assessing disparity in employment decisions. In broad terms, these perspectives can be categorized into two general schools of thought: statistical significance testing and practical significance measurement. Statistical significance testing usually takes the form of null hypothesis significance testing, and evaluates the probability that a disparity is due to chance. In general, disparities that are not due to chance will increase the confidence of scientists, practitioners, and decision-makers that "real" differences exist. Practical significance, however, focuses on whether a disparity is of a magnitude that is meaningful; in other

words, is a disparity large enough for society, an organization, and/or a decision-maker to notice? In many situations, this analysis considers whether the magnitude is large enough for a layperson (i.e., not a lawyer or statistical expert) to conclude that it is meaningful. Given this scenario, we asked TAC members a variety of high level questions, including:

- What types of analytic methods are most useful for assessing a difference in employment decision rates?
- Which methods of assessment would you generally recommend when assessing adverse impact?
- How confident would you be in concluding meaningful adverse impact across a variety of scenarios?

Table 3.1 presents TAC member ratings of the usefulness of various adverse impact measurement methods.[6] As the data show, statistical significance tests were rated as most useful; in fact, 96% thought that these measures were either somewhat or very useful. This is consistent with the broad set of case law that has endorsed statistical significance testing since *Hazelwood School District vs. United States* (1977). Interestingly, the 4/5th rule, which is a decision rule (e.g., substantial / not substantial) applied to the impact ratio that is endorsed by the UGESP, had the highest percentage of members that rated the measure as not useful (27%). This notion is somewhat consistent with some older (e.g., Boardman, 1979, Greenberg, 1979) and recent (e.g. Roth et al., 2006) literature suggesting that the 4/5th rule has some counter-intuitive psychometric properties, and may too often flag false positives or not flag true differences (i.e., false negatives). Other effect size measures (e.g., odds ratios) that are intended to assess the magnitude of disparity but without specific decision rules were endorsed as at least somewhat useful by 80% of respondents. Other types of practical significance measures intended to assess the stability of results in non-technical terms (e.g., flip-flop rules such as those found in *Waisome vs. Port Authority*, 1991 and *Contreras vs. City of Los Angeles,* 1981) were

[6] For a basic review of these measurement methods, we suggest reading Meier, Sacks, & Zabell, (1984), Biddle (2005), and Collins and Morris (2008).

endorsed as at least somewhat useful by 74% of respondents. Between 45 and 52 respondents answered these questions.

When asked which measures TAC members typically use, 26% responded that they only use statistical significance tests; 30% used a combination of statistical significance tests and the 4/5th rule; and 26% used a combination of statistical significance testing, the 4/5th rule, and another practical significance measure. Eighteen percent of survey respondents responded "other" to this question, and open ended responses suggested some form of multi-measure combination. Interestingly, no member endorsed using the 4/5th rule alone. Overall, 50 TAC members responded to this question.

Table 3.1: What methods of assessment are most useful for assessing a difference in employment decision rates?

Method of Assessment	Not Useful	Somewhat Useful	Very Useful	Response Count
Statistical significance Tests	4.00%	38.00%	58.00%	50
4/5th rule	26.90%	53.80%	19.20%	52
Other effect size measures (e.g., odds ratios)	19.60%	50.00%	30.40%	46
Practical significance measures (e.g., UGESP flip- flop rule)	25.50%	55.30%	19.10%	47

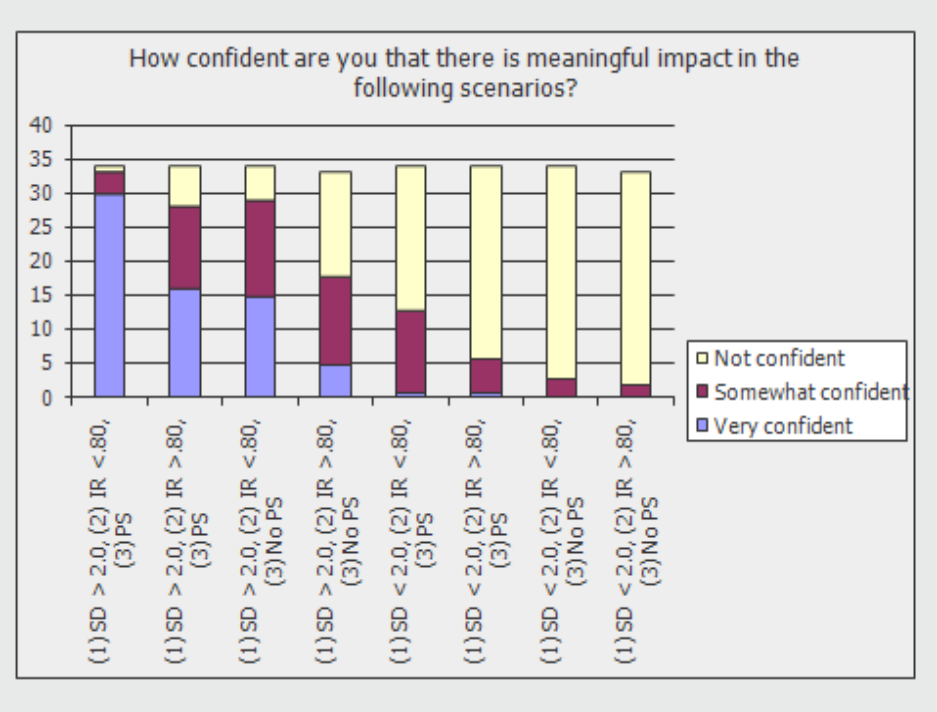
Figure 3.1 shows TAC member confidence when interpreting results from various adverse impact measurement methods[7]. Overall, 35 TAC members answered this question. Results demonstrate what will become a recurring theme in this chapter; TAC members are generally more confident in results when multiple measures are used. This result is consistent with the notion of a "preponderance of evidence" standard. Focus group discussions reiterated similar themes, such that impact should be viewed along a continuum of sorts as opposed to a "present/absent" dichotomy. In this context, multiple measures may be used, and the more measures that suggest disparity, the more confident experts may be in making overall conclusions concerning meaningfulness. Of particular

[7] SD = statistical significance test; IR = Impact Ratio; PS = other measure of practical significance.

importance is the need to look at both the magnitude of the disparity in outcomes (through some measure of effect size or other practical significance measure) as well as the likelihood of that result occurring due to chance (utilizing statistical significance tests). Using only one of these strategies may ignore important information and lead to incorrect conclusions.

It should also be noted, however, that there is often considerable redundancy among methods within paradigms. For example, over 20 statistical significance tests are available for the analysis of 2-by-2 tables, and in the vast majority of situations these will provide the same conclusion. Alternate measures of effect size (e.g., impact ratio, odds ratio, selection rate difference) are all related. Similarly, shortfall analysis is correlated with the sample size and selection rate difference, and thus may provide redundant information. Care should be taken to avoid artificially inflating perceived confidence by aggregating largely redundant statistics. That being said, there are some narrow circumstances (e.g., very high or low selection rates) where different effect sizes may provide different conclusions.

Figure 3.1: TAC member confidence in interpreting results from multiple adverse impact measurement methods



However, based on focus group discussion, there was substantial disagreement over which measurement strategies should be used. Some members argued that statistical significance testing should be the primary measurement tool to assess disparity, whereas others in the focus group argued for the use of multiple methods, particularly when sample sizes are very large or very small. The following sub-sections of this chapter deal with each of these measurement methods in more detail.

Recommendation 3.1: *Multiple measurement methods are available to assess adverse impact. Generally, the more measures that are used (whether those are within a particular paradigm or across paradigms), the more confident analysts, lawyers, and decision makers can be in judging the meaningfulness of a disparity. Multiple methods of adverse impact detection should be used. When using multiple methods, however, care should be taken to minimize redundancy and combine only methods that each add unique information.*

Statistical Significance Testing

Several survey items asked TAC members about statistical significance testing. One issue that caused substantial disagreement among TAC members concerned what the most appropriate statistical significance model (e.g., some form of the binomial or hyper-geometric model[8]) is for various data structures.[9] In this context, the issue was whether

[8] See Gastwirth (1988) for a concise review of these two discrete probability models. Binomial models assume sampling with replacement, such that the probability of each employment decision is the same. The hypergeometric distribution, on the other hand, assumes sampling without replacement, such that the probability of the first selection from a pool will be slightly different than the second because there is one less candidate in the pool. In most situations these probability distributions will produce similar results, but when sample sizes or expected values are small, results may vary meaningfully. Some selection systems (e.g., hiring applicants from a pool, terminating a number of employees from a workforce, making pass-fail decisions on a test) may be more appropriately mirrored by one probability model than another.

[9] It should be noted that the presentation here greatly simplifies the topic. This topic has been debated by statisticians for decades, and involves subtle philosophical arguments that are not (and for practical reasons could not be) fully represented here. Some TAC members argued that it is not a simple matter of mapping situations onto models, and that in some situations there may not be a "most appropriate model" given an ambiguous selection data structure. However, the majority of TAC members felt strongly that the independence of selection decisions should drive the model.

subgroup representation (i.e., the percentage of subgroup members in the applicant pool) and the number of employment decisions were known beforehand. Certain statistical significance probability models assume that this information is or is not known, and this assumption affects the probability of employment decisions.

This issue also subsumed a question included on the TAC survey: Should data analyzed in the EEO context (e.g., Title VII litigation, OFCCP audits) be considered "fixed" because it is retrospective in nature? Interestingly, 58% of survey respondents answered "yes" when asked whether data in the EEO context should be considered fixed because it is retrospective in nature. Overall, 33 TAC members answered this question. However, the majority of focus group members strongly disagreed with the majority of survey respondents. In addition, focus group members were concerned that this survey question may not have been clear enough and that results could be misleading if respondents did not understand the question. Given this context, it is difficult to make any firm conclusions about this issue.

Some focus group members suggested that this may be a legal argument more than a statistical one and that they have worked with lawyers who argue that all EEO data are fixed, and a hypergeometric model would best mirror the reality of the selection. While some members advocated this retrospective notion the majority of focus group members disagreed and argued that it was the selection model (i.e., how the data were collected in reality) that dictates which statistical model is most appropriate, not the fact that the data are retrospective. Focus group members did note that in many situations, it is unclear how the decision making process under scrutiny functioned in reality, and that in such a case it is reasonable to consider multiple models.

Members also differentiated a litigation model from a proactive analysis, and suggested that this could play a role in support for one model over another (e.g., proactive analysis of an incomplete hiring cycle may not truly be fixed). Table 3.2 shows survey results capturing the appropriateness of statistical models for various data structures. Overall, 26 TAC members answered this survey question. These examples were intended to represent

some examples on a continuum ranging from clearly knowing the number of employment decisions and subgroup representation to clearly not knowing the number of employment decisions and subgroup representation. As described earlier, focus group members suggested interpreting these results with caution given the difficulty and "it depends" nature of the question.

Data analysis in a reduction of force context was used as an example where the hypergeometric model may be most appropriate, because subgroup representation is usually known (e.g., the current set of employees), as is the number of employment decisions to be made (e.g., a need to reduce the workforce by five). However, if other criteria are used (e.g., reducing a fixed payroll by a certain percentage) instead of the actual number of selections, the statistical model should account for this when possible.

Data analysis of pass/fail results from applicant flow hiring data was used as an example where neither the subgroup representation nor the number of employment decisions is fixed, and thus a traditional double binomial model may be more appropriate. Analyses of promotions were discussed and TAC members suggested that these analyses could be ambiguous, because there may or may not be readily available applicant pools, the number of promotions may or may not be made beforehand, and some promotions may not be appropriate to analyze together (e.g., competitive vs. career progression). Some TAC members noted that, in the vast majority of situations, results from various models will be similar, and choice of one specific model over another may not have meaningful consequences.

Table 3.2: What is the most appropriate statistical model (binomial vs. hypergeometric) for various data structures?

Answer Options	1 - Binomial	2 - Hypergeometric	Response Count
Hiring a fixed number of persons from an already existing applicant pool	35.00%	65.00%	26
Hiring a fixed number of persons from an applicant pool that has not yet been defined	71.00%	29.00%	21
Hiring an unknown number of persons from an applicant pool that has not yet been defined	76.00%	24.00%	17
Hiring an unknown number of persons from an already existing applicant pool	47.00%	53.00%	17
Making pass/fail decisions at a predetermined cut score from an already existing applicant pool	58.00%	42.00%	24
Making pass/fail decisions at a predetermined cut score from an applicant pool that has not yet been defined	81.00%	19.00%	22
Making terminated/not terminated decisions from a predetermined employee list	41.00%	59.00%	22

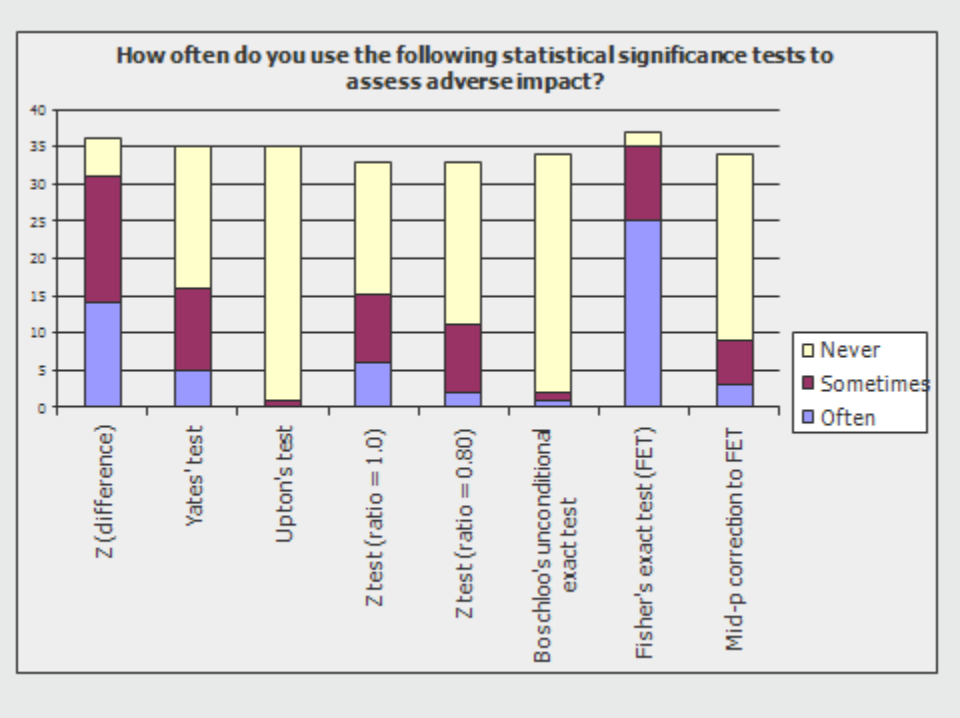
Recommendation 3.2: *It is important to determine the context in which the data are being analyzed. Substantial effort should be made to understand the reality of the employment decision process under scrutiny and to mirror that process in analyses. The statistical model (e.g., some form of binomial or hypergeometric) that best mirrors the reality of the employment decision process should be used. It is often difficult to understand how employment decisions were made, and thus challenging to identify the most appropriate model.*

Another set of survey questions focused on which statistical significance tests are used most often in practice, and which tests TAC members opine are most appropriate. Figures 3.2 and 3.3 present survey results for a variety of statistical significance tests that may be used in the adverse impact context. Although this list isn't exhaustive, it was intended to include tests commonly seen in EEO contexts. Overall, 37 TAC members answered the statistical significance test "use" survey question, whereas 34 members answered the statistical significance test "appropriateness" question.

Once again, focus group members urged caution in interpreting these results for a variety of reasons, most notably because the frequency and appropriateness of particular statistical significance tests depends on how the employment decisions were made in reality. Thus, if someone has worked on more cases where hiring processes were fixed, more hypergeometric models would likely be used and deemed appropriate. If an expert worked on more cases where the number of employment decisions and/or subgroup representation were unknown, more binomial models would likely be used and deemed appropriate Additionally, the focus group discussed the notion that in some cases, certain tests are used simply because experts think (or know) that opponent experts or agencies will run those tests. For example, OFCCP usually runs a Z-test based on the double binomial model (unless sample sizes are small, when a Fisher's exact test is conducted), regardless of how the employment decision system functioned in practice.

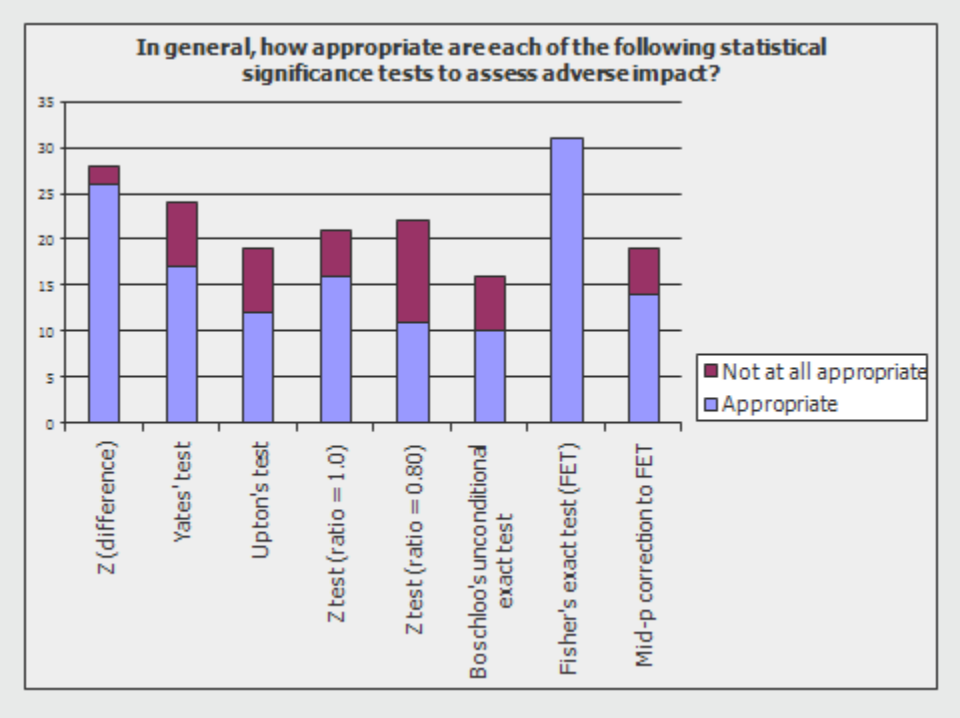
As survey results show, Fisher's exact test was generally the most supported hypergeometric test; over 90% of respondents used it at least sometimes, endorsed it as appropriate, and selected it as most appropriate from the list of tests. The double binomial Z-test was the most supported binomial model; over 80% of survey respondents used it at least sometimes, over 90% thought it was appropriate, and about 40% endorsed it as most appropriate. However, the majority of focus group members would not endorse one particular binomial test as preferred, or one particular hypergeometric test as preferred. Instead, members reiterated that context matters, and that more than one test may be reasonable once the appropriate statistical model has been identified to the extent possible.

Figure 3.2: Frequency of statistical significance test use



It is interesting to note that the corrections to Fisher's exact test such as Lancaster's mid-p were used by about one-third of respondents, were deemed appropriate, and were endorsed as most appropriate by about 10% of respondents. This correction has received some recent attention in the biomedical field (Crans & Shuster, 2008; Lin & Yang, 2009), and has just recently been mentioned in the EEO community at the time this report was written. Some research has suggested that the correction produced Type I error rates that are closer to the nominal alpha level than the uncorrected Fisher's test, which some members suggested was overly conservative when sample size is small. However, the mid-p does not appear to be in widespread use in the EEO community, and focus group members intentionally chose not to discuss it in detail. It will be interesting to see if the mid-p correction becomes more frequently used and is perceived as appropriate in the next few years as part of disparity analyses. We suggest monitoring the scholarly literature on this topic.

Figure 3.3: Appropriateness of statistical significance test use



Recommendation 3.3: *Multiple statistical significance tests may be appropriate in a given context. Care should be taken to use tests that employ the most appropriate statistical significance model. The TAC does not endorse particular tests as preferred under various probability models.*

Another set of survey questions focused on some specifics of statistical significance testing. For example, one question asked what alpha level, or probability standard, was most appropriate for adverse impact analyses. The p-value, which is compared to the alpha level, is the probability that a particular outcome could have occurred by chance. When a probability value is 1.0, researchers are essentially 100% confident that a finding occurred by chance. If a probability value is 0.05, a researcher is generally confident that a finding would have occurred by chance only once out of twenty times.

Overall, 41 TAC members answered this survey question. Eighty-eight percent of survey respondents endorsed an alpha level of 0.05, which corresponds to a standard deviation value of about 2.0 using a two-tailed test. The other 12% of respondents replied that "it

depends". Open ended comments for this group suggested that some members prefer a scenario where alpha levels depend on sample size. For example, larger sample sizes may warrant lower alpha levels, while smaller sample sizes may warrant higher alpha levels. Although the majority of the focus group advocated this idea, no participant had seen this perspective actually used in court.

Interestingly, no respondent endorsed alpha levels of 0.01 (about three standard deviations for a two-tailed test) or 0.10 (about 1.65 standard deviations using a two-tailed test). Focus group discussion reiterated the notion that both professional (e.g., standards for scholarly journals) and case law evidence support an alpha level of 0.05, although some case law is ambiguous (e.g., the two or three standard deviation criterion endorsed by the Supreme Court in *Hazelwood School District vs. United States*[10],1977).

Another survey question focused on whether one-tailed or two-tailed significance tests should be used. Of the 45 TAC members answering this question, 76% endorsed two-tailed tests, whereas 24% endorsed one-tailed tests. Focus group discussion generally supported the use of a two-tailed test, although there was some disagreement. Some members who advocated the use of one-tailed tests suggested that, from a scientific hypothesis testing perspective, a particular direction of disparity was hypothesized. In other words, if a claim of discrimination is made, it is implied that the claimant group is impacted. Those in support of two-tailed tests countered that analyses in the EEO context occur under statutes and mandates where all groups are protected (e.g., Title VII, Executive Order 11246), and thus experts and decision-makers are interested in disparities that occur against any group.

One other interesting issue concerned situations where disparities in employment test results are at the center of the case. In these situations, a long line of research may suggest that the disparity will occur against one group. For example, subgroup

[10] Interestingly, the Hazelwood case did not include analysis of applicant flow data, and was actually a pattern or practice allegation of discrimination. In that case the standard deviation criterion was used to evaluate a utilization analysis comparing the percentage of teachers that were hired to the available percentage of teachers in the labor pool.

differences in cognitive ability test scores often exist in favor of White test takers as compared with Black and Hispanic test takers. Likewise, subgroup differences in physical tests often exist in favor of men as compared with women. In these situations previous research may support a directional (e.g. one- tail test) hypothesis.

However, other focus group participants countered that the law is still the law, that everyone is protected equally under the 14th or 5th amendments of the Constitution, and that in practice a test could produce counter-intuitive results for a variety of reasons. Some members reiterated that, in the end, the difference between a one-tailed or a two-tailed significance test may be determined by the courts, and that it is important to consider legal guidance for determining the accepted statistical criteria when there really isn't a right answer from a scientific perspective.

Some focus group members mentioned that, in many proactive analysis situations, there is not a claim of discrimination that could "fix" a tail. Here an analysis comparing all possible groups to all other groups is possible. Sometimes the highest selected or passing group is compared to all other groups. In other cases, traditionally advantaged and disadvantaged groups are compared. This is also the case for many present day OFCCP audits, where no claim of discrimination is made, and the agency may conduct any number of disparity analyses.

Other focus group members noted that the UGESP aren't particularly informative on the issue of statistical significance. In fact, the only section of UGESP that covers this issue with any specificity is in the context of assessing the statistical relation between a selection procedure and an outcome in a validation study. Specifically, the UGESP state*: Generally, a selection procedure is considered related to the criterion, for the purposes of these guidelines, when the relationship between performance on the procedure and performance on the criterion measure is statistically significant at the 0.05 level of significance, which means that it is sufficiently high as to have a probability of no more than one (1) in twenty (20) to have occurred by chance.*

The statement above implies a two-tailed test at the 0.05 alpha level. Focus group members did agree that a 0.05 alpha level should be used for two-tailed tests, and a 0.025 alpha level should be used for one-tailed tests. Focus group participants agreed that it is useful to transform probability values into standard deviations, so that decision makers can better understand the information. However, some members cautioned that the standard deviation can be misinterpreted as a practical significance measure of magnitude, which it is not. It is a metric that allows for a decision rule to be made on whole numbers (e.g., two standard deviations or greater is statistically significant), but does not allow for useful inferences related to magnitude. For example, a disparity of five standard deviations is not necessarily more 'substantial' adverse impact than is a disparity of three standard deviations; in both instances we are very confident that the difference in rates is not due to chance. However, this metric does not inform on magnitude and the disparity more likely to be due to chance may be more substantial using other metrics

Recommendation 3.4: *This committee generally recommends that using a two standard deviation criterion is usually most appropriate to determine statistical significance. Determining whether to use a one-tailed or a two-tailed test will depend upon the context under which these analyses are being performed (litigation or a proactive environment). However, a 0.05 alpha level should be used in the two-tailed case and a 0.025 alpha level should be used in the one-tail case.*

Another issue of discussion concerned whether exact tests should be preferred over approximation tests that are based on large sample theory. Exact tests do not require large sample or cell sizes for accurate probabilities. Based on survey results of 35 TAC members, 48% of respondents always preferred exact tests, whereas about 25% preferred exact tests only when sample sizes are small (< 30), and another 17% of respondents preferred exact tests when sample sizes were small and expected cell sizes were also small (e.g., less than five). Focus group members did not have a strong opinion on this issue and generally agreed that exact tests should be used when feasible. Others reiterated that the appropriate model was a more important consideration, and that exact tests and estimator tests will provide the same conclusion in the vast majority of cases. Other

participants pointed out that software is now available that easily handles large sample sizes (e.g., StatExact, LogExact), such that exact tests can be computed conveniently in just about all situations. Focus group members also noted that the term exact is not a synonym for most accurate, and that exact tests can be used inappropriately.

Recommendation 3.5: *The committee suggests that exact statistical significance tests (e.g., Boschloo's unconditional exact test, some form of Fisher's exact test) should be used when possible. The specific test depends on which probability model is most appropriate, and in many instances exact tests and estimator tests under the same probability model will result in identical conclusions*

One issue that produced substantial disagreement among experts concerned correcting the alpha level when multiple comparisons are made across groups. Specifically, experts were asked whether they would recommend that the statistical significance criterion be corrected to account for the number of statistical tests. In other words, should the standard deviation criterion (e.g., two or more) be adjusted upward to account for the fact that multiple tests are being conducted (e.g., multiple subgroup analyses, analyses of multiple jobs, across multiple locations). For example, assume five tests are conducted on the same job. A traditional two-tailed alpha level of 0.05 (or about two standard deviations) could be corrected by the number of tests (0.05/5 comparisons), which is an alpha of 0.01 (or about 2.5 standard deviations). Of the 28 TAC members answering this survey question, over 57% reported that the significance criterion should be corrected for the number of statistical tests conducted.

Once again, focus group members reiterated that this decision depends on the context and whether it is a proactive or reactive analysis. For example, each separate claim of discrimination (i.e., a reactive analysis where the research question is specified by the claim) may warrant separate consideration and alpha levels. However, in a proactive situation such as an OFCCP audit, any number of comparisons could be made (highest selected vs. subgroups, traditional advantaged vs. disadvantaged, total minority vs. non-minority, all pair-wise comparisons, etc.), and correcting alpha levels for multiple

comparisons within research question (e.g., all analyses within a job, all analyses comparing different racial/ethnicity groups) may be reasonable. In these latter cases the actual alpha within research question may be substantially greater than p <= 0.05 (or correspondingly, to a standard deviation substantially smaller than 1.96). Most members suggested a Bonferroni correction, which adjusts the alpha level based on the number of comparisons made.

However, at least one focus group participant cited literature that the Bonferroni correction may be too conservative, and other methods should be considered. Other participants pointed out that in most litigation, this is a non-issue because the claim defines a single analysis, and noted that they did not adjust alpha in litigation. Other participants cited an OFCCP statistical panel report from 1979, which considered multiple comparisons and suggested controlling for multiple comparisons. Participants noted that, in this situation, all possible comparisons were possible. However, a policy decision will necessarily require guidance on when to correct and not to correct alpha, and how to ensure that corrections are not abused. It is also important to note that this issue may be very difficult to explain to decision makers.

Recommendation 3.6: *There is general disagreement on the issue of correcting alpha for multiple comparisons. However, committee members generally agree that the potential for capitalizing on chance should be considered in all analyses, perhaps through a more descriptive approach (e.g., reporting at multiple alpha levels, showing the total number of analyses conducted).*

A few other issues related to statistical significance came up during focus group discussion and are worth noting. First, the issue of whether it would be useful to report statistical power was considered, particularly in contexts where either statistical power was very low (e.g., close to zero) due to small sample sizes (e.g., less than 20) or when statistical power was very high (e.g., close to one) when sample sizes were very large (e.g., 10,000).

Focus group members agreed that statistical power information would be useful in giving weight to the results of a statistical significance test (i.e., less weight in very small or very large samples where statistical power was very low or very high). However, members also suggested that the concept of statistical power is highly technical, would be very difficult to explain to decision makers, and may confuse more than aid in understanding whether a disparity is meaningful. In situations where samples are small, a flip-flop rule may convey similar information in a more parsimonious fashion. Likewise, in situations where sample sizes are very large, practical significance measures may convey much more useful information in a more parsimonious fashion.

Another question asked by focus group members concerned how experts handled tables where there were either more than two groups (e.g., Whites, Hispanics, and Blacks) or where employment decisions were trichotomous instead of dichotomous (e.g., promoted, promoted with distinction, and not promoted). In this situation, some members recommended reporting an omnibus Chi-Square test, and if there is overall disparity, sub-analyses can be organized into separate 2-by-2 tables. Members reported that this method has been used in OFCCP audits and Title VII litigation. Another possibility is to use a statistical model that explicitly includes selections from more than two groups, such as the multinomial, although the analyst may need to conduct other analyses to identity which specific group comparisons drive the multinomial results.

Practical Significance

Another issue on which TAC members spent substantial time concerned practical significance measures of disparity. In contrast to statistical significance, practical significance tests assess the magnitude of a disparity via metrics that are intended to be understandable to the layperson and do not require a background in statistical significance testing. As stated by UGESP, "*the Guidelines do not rely primarily upon a test of statistical significance, but use the 4/5th rule of thumb as a practical and easy-to-administer measure of whether differences in selection rates are substantial".*

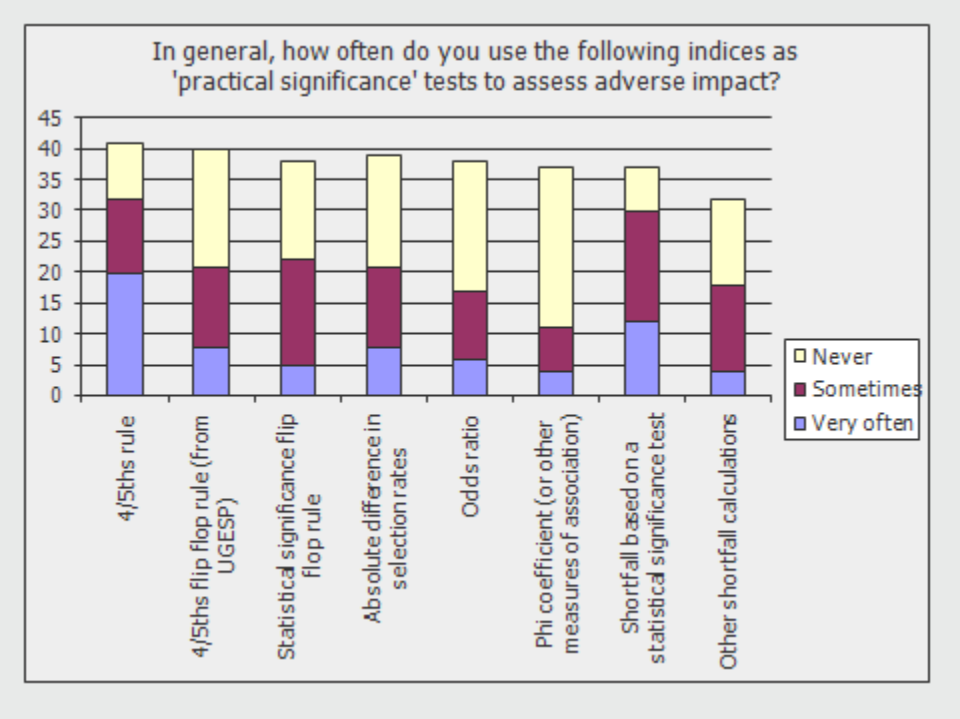
It is important to distinguish between the related concepts of effect sizes and practical significance rules. An effect size is a measure of the magnitude of the difference in selection rates between groups. Common effect sizes for adverse impact data include the adverse impact ratio, the selection rate difference and the odds ratio. Effect sizes exist on a continuum reflecting the magnitude of the difference between groups. A practical significance rule is the application of some decision criterion to an effect size. The most common of these is the 4/5th rule, which flags a decision as practically significant if the adverse impact ratio is less than 0.80. There are other "flip-flop" measures of practical significance that do not assess magnitude, but instead assess the stability of results. TAC members noted that these are generally redundant with the purpose of statistical significance testing.

Some of the more general topics in this section concerned which 'practical significance' measures were used, and which were considered adequate. Figures 3.4 and 3.5 present the survey results of the 41 TAC members answering these questions. As Figure 3.4 shows, the 4/5th rule was used most frequently by respondents and considered appropriate by a majority of respondents. Note that TAC members in the focus group strongly disagreed on this point. In fact, more than half of focus group participants did not use the 4/5th rule at all, and suggested that it was arbitrary, potentially misleading, and exhibited poor psychometric properties. It should be noted that these criticisms focused on the decision rule (i.e., the 4/5th rule standard) more than the effect size (i.e., the impact ratio). This group generally preferred statistical significance tests to practical significance measures.

However, the other focus group participants disagreed, and although there was some agreement concerning its poor psychometric properties, still used the 4/5th rule because it was UGESP endorsed and is still used often in EEO contexts. For example, the Supreme Court considered it as recently as 2009 (*Ricci vs. DeStefano*). It should be noted that this group of participants generally had more positive perceptions regarding practical significance measures. About 60% of survey respondents suggested that the 4/5th rule was the most appropriate practical significance measure. This led to some very interesting

debate about the current state of selection, how the internet has created applicant pools that are very different from those that existed in the 1970s, and the pros and cons of different adverse impact measurement perspectives. For example, some participants suggested that practical significance measures are critical in situations where statistical significance testing is less useful (e.g., very large or small sample sizes).

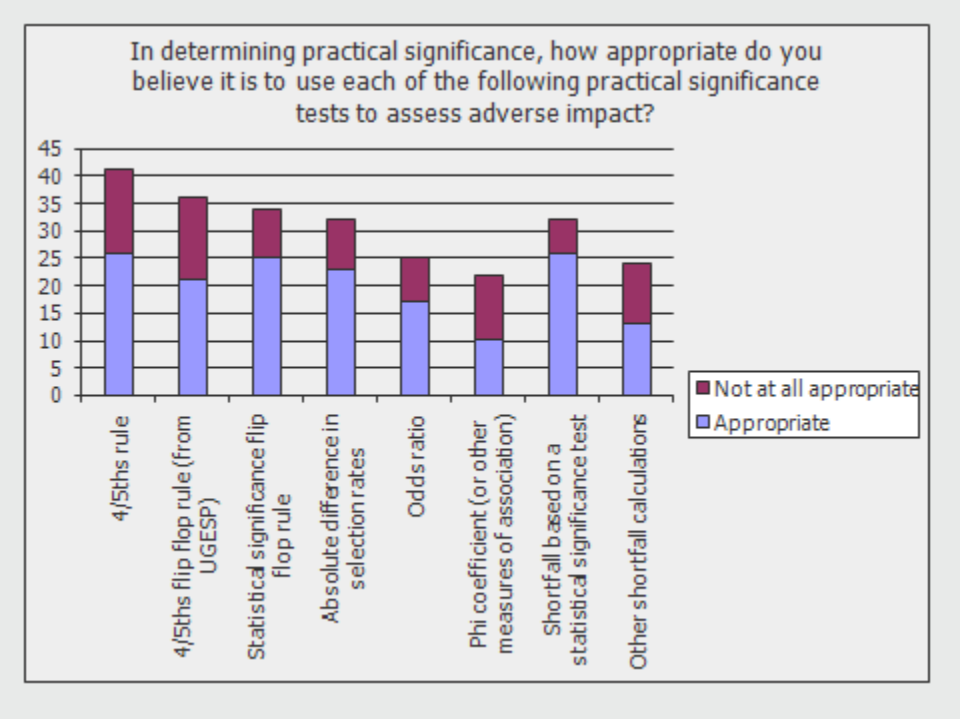
Figure 3.4: Frequency of practical significance test use



It is interesting to note that shortfalls (the difference between expected and observed number of employment decisions at the heart of statistical significance tests) and flip-flop rules (either UGESP endorsed regarding the highest selected group, or case law endorsed such as the "*Waisome* flip-flop" rule to non-significant statistical disparity) were used at least sometimes by a majority of respondents as measures of practical significance. Some focus group members suggested that flip-flop rules are useful only in small sample situations, and are generally redundant with the purpose of statistical significance tests (to assess the confidence or stability of results).

The use of a shortfall as a practical significance measure produced some interesting discussion.[11] Specifically, it was noted that the shortfall is dependent on sample size much like the probability (or standard deviation) value of a statistical significance test. In this scenario the shortfall may be somewhat trivial and difficult to interpret when sample sizes are very small or very large. One suggestion that was well received by the focus group was to consider the shortfall relative to either the total number of applicants or the total number of disadvantaged group applicants, and use that ratio as a measure of practical significance.

Figure 3.5: Appropriateness of practical significance tests



With regard to other potential measures of practical significance, focus group members noted that the raw percentage difference can be useful as a practical significance measure. This has been endorsed in case law (e.g., *Moore vs. Southwestern Bell*, 1979; *Frazier vs. Garrison*, 1993), where differences of 7% and 4% were deemed not practically

[11] It should be noted that the legal usefulness of the shortfall is discussed in the legal/policy chapter (IV) with emphasis on the magnitude of liability.

meaningful.[12] It was also interesting to note that other potential measures such as odds ratios, correlation coefficients, and other shortfall calculations were not used often or considered appropriate by the vast majority. Focus group discussion provided more insight on this issue, and the explanation may be that these measures won't provide substantially different information than more established measures in the vast majority of cases, and are more difficult to explain to decision makers.

In terms of rules of thumb regardless of the index used, the focus group participants described practical significance standards as a sliding scale. Once again, it is important to note that some focus group participants suggested that practical significance measures were less valuable than statistical significance tests. The focus group also discussed some basic practical significance information from the UGESP Questions and Answers (1979), including:

- <u>The use of multiple measures at once (Question 20):</u> *If, for the sake of illustration, we assume that nationwide statistics show that use of an arrest record would disqualify 10% of all Hispanic persons but only 4% of all whites other than Hispanic (hereafter non-Hispanic), the selection rate for that selection procedure is 90% for Hispanics and 96% for non-Hispanics. Therefore, the 4/5th rule of thumb would not indicate the presence of adverse impact (90% is approximately 94% of 96%). But in this example, the information is based upon nationwide statistics, and the sample is large enough to yield statistically significant results, and the difference (**Hispanics are two and a half times as likely to be disqualified as non-Hispanics) is large enough to be practically significant.***

- <u>The use of a shortfall to exemplify practical significance (Question 20):</u> *On the other hand, a difference of more than 20% in rates of selection may not provide a basis for finding adverse impact if the number of persons selected is very small.*

[12] Note that it is reasonable to expect that the relevance of these types of rules of thumb depend where on the selection procedure continuum results fall. In the cases cited, selection rates were very high (e.g., higher than 85% for each group) and the 4/5th rule was not violated. Interpretations may change if a 4% selection rate is compared to a 0% selection rate.

For example, if the employer ***selected three males and one female from an applicant pool of 20 males and 10 females,*** *the 4/5th rule would indicate adverse impact (selection rate for women is 10%; for men 15%; 10/15 or 66 2/3% is less than 80%),* ***yet the number of selections is too small to warrant a determination of adverse impact.***

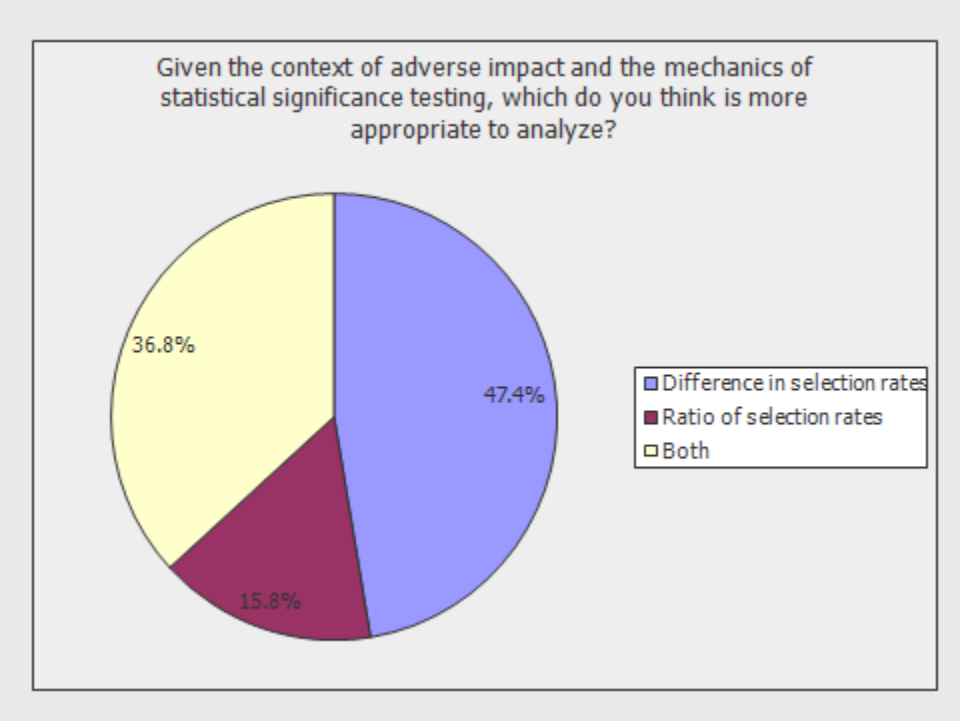
- The use of a flip-flop rule to exemplify practical significance (Question 21): *If the numbers of persons and the difference in selection rates are so small that it is likely that the difference could have occurred by chance, the Federal agencies will not assume the existence of adverse impact, in the absence of other evidence. In this example, the difference in selection rates is too small, given the small number of black applicants, to constitute adverse impact in the absence of other information (see Section 4D). If only one more black had been hired instead of a white the selection rate for blacks (20%) would be higher than that for whites (18.7%).*

Recommendation 3.7: *Practical significance should be considered and a variety of effect sizes (e.g., impact ratio, odds ratio, rate difference) may be useful measures. However, specific decision rules (e.g., 4/5th rule) for interpreting these effect sizes were deemed arbitrary and potentially misleading. Flip-flop rules are less useful, but may be informative when sample size is small. Practical significance measures should be paired with a statistical significance test.*

Another question asked of TAC members focused on what type of information (or effect size) should be used in the analysis of disparity. These analyses inherently require a comparison of information across two groups. In other words, should the EEO community focus on differences in selection rates, or some form of ratio of those rates (e.g., the impact ratio used for 4/5th rule analysis, an odds ratio that includes information about both positive and negative rates)? Note that this decision also has implications for statistical significance testing as well, although in large sample sizes results may be identical (Morris & Lobsenz, 2000).

Figure 3.6 shows survey results for this question. Forty-seven percent of the 38 survey respondents endorsed using a difference in employment decision rates, 16% endorsed using a ratio of rates, and 37% suggested using both. Focus group participants suggested that, as long as the statistical model being used is appropriate, this question is more about framing the presentation more than anything else, and chose not to make a formal recommendation.

Figure 3.6: Which effect sizes should be measured in an adverse impact analysis



Another survey question specifically asked if it was appropriate to use a 4/5th rule analysis of rejection or failure rates instead of selection or pass rates (note that it makes no difference for statistical significance tests). Seventy-one percent of the 31 survey respondents said it is never appropriate to use rejection rates. Of those who said it may be appropriate in some cases, comments suggested this method may be appropriate for the analysis of termination and reduction in force decisions, where a decision to terminate/lay off is the primary organizational activity of interest.

This issue was echoed by focus group participants, who once again noted the problems with the 4/5th rule, and suggested that an analysis of termination rates may require flipping the group numerator/denominator to assess what percentage of the disadvantaged group termination rate is relative to the advantaged group termination rate (instead of a disadvantaged group numerator and advantaged group denominator).

However, there was general agreement that when the analysis focused on such traditional positive employment decisions as hiring, promotion, or passing a test, the impact ratio should be computed and 4/5th rule standard applied using selection rates, if for no other reason than to maintain consistency. In other words, a 4/5th rule analysis of rejection rates may be more confusing than useful. Members pointed out that using selection rates in calculations is consistent with the examples from UGESP, and cited Gastwirth (1988) for a more detailed treatment where the same general recommendation was made.

Recommendation 3.8: *When an analysis focuses on traditional positive employment decisions such as hiring, promotion or passing a test, the impact ratio should be computed using selection rates, not rejection rates. Analysts should consider context and the type of employment decision being made to identify caveats to this recommendation (e.g., analyses of termination decisions).*

Another question focused on a "hybrid" measure of disparity that combines statistical and practical significance perspectives via computing confidence intervals around such effect size measures as the impact ratio and the odds ratio. In this scenario, a point estimate would indicate the best guess for the actual effect size, and confidence intervals would show the precision of the estimate as well as whether or not that estimate is statistically significant (e.g., contains zero or one in the interval depending on the effect size).

Ninety-one percent of the 35 survey respondents said that confidence intervals were somewhat or very useful. However, there was strong disagreement over this issue in focus group discussion. Some participants suggested that it would be too difficult to explain confidence intervals to decision-makers, and that these intervals do not provide

new information relative to separate statistical significance tests framed as standard deviations and practical significance measures. Similarly, this group of participants suggested that the danger of misinterpretation would outweigh the value of correctly interpreting the intervals. Others in the focus group disagreed, suggesting that intervals were useful as a practical measure to the layperson, because they provide a more intuitive presentation of the precision of the analysis and thus the confidence of results.

Given the general disagreement on topics subsumed within practical significance measurement, the TAC cannot make a clear recommendation on this issue other than practical significance should be considered. Perhaps the OFCCP Statistical Standards Panel Report (1979) guidance on practical significance is still applicable today: "*any standard of practical significance is arbitrary. It is not rooted in mathematics or statistics, but is a practical judgments to the size of the disparity….Second, no single mathematical measure of practical significance will be suitable to widely different factual settings*…" Thus, context should be considered when deciding what practical measures to use and on what decision rules to base conclusions.

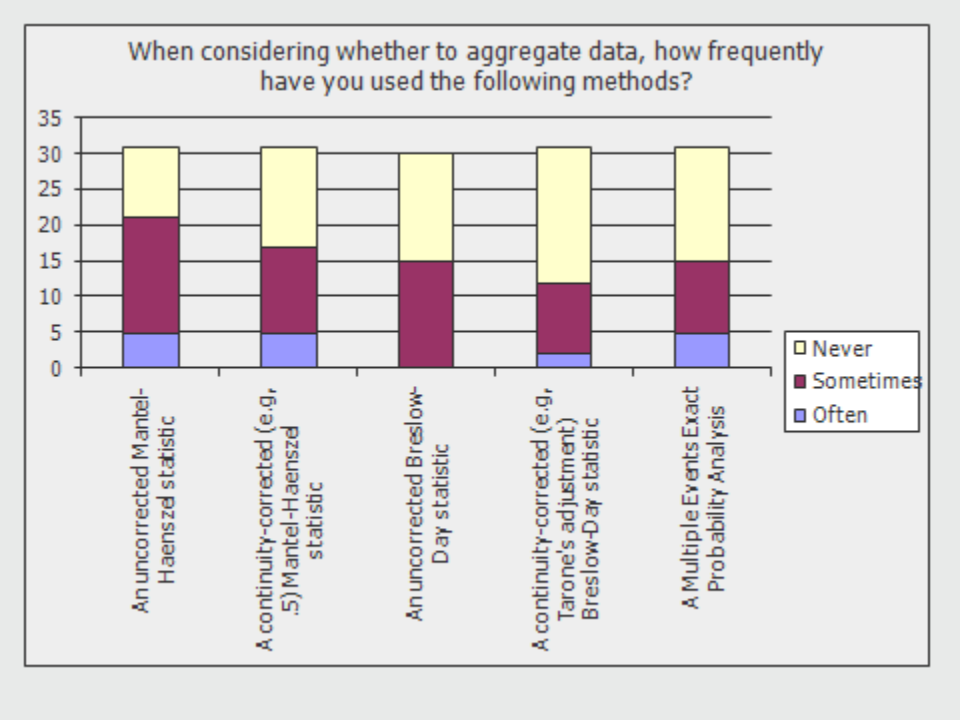
Data Aggregation

Another issue that is often contentious in the legal realm is how to appropriately aggregate or disaggregate data. For example, an adverse impact analysis may focus on multiple jobs, span across multiple locations and time periods, or include multiple protected groups. Given this context, the analysis could be conducted in a number of different ways (e.g., combining across strata into a single pool, conducting analyses separately across different levels, using more complex weighting methods). How the data are aggregated can make a meaningful difference in the results of the analysis, and once again, it is reasonable to assume that mirroring the reality of the employment decision process is a goal. Toward that end, CCE asked a number of questions related to data aggregation and the statistical methods used to determine both whether aggregation is appropriate and appropriate analytic aggregation methods.

When asked about the types of strata (i.e., levels of analysis) members felt comfortable aggregating, survey comments suggested that the level of TAC member comfort depended on a variety of factors, and that it may be reasonable to aggregate across any strata (i.e. a so-called "multiple pools" analysis) if that is the reality of the employment decision process that was used. For this reason, survey results were difficult to interpret. In focus groups, discussion once again centered around the idea of mirroring reality, and using both statistical and theoretical considerations to drive the analysis. For example, if the same employment decision process is used across multiple locations, multiple jobs, or multiple time periods, it may be reasonable to aggregate across these strata. Likewise, as the next few topics show, there are statistical methods used by TAC members to assess whether employment decisions are similar across various strata, and that these methods are well accepted.

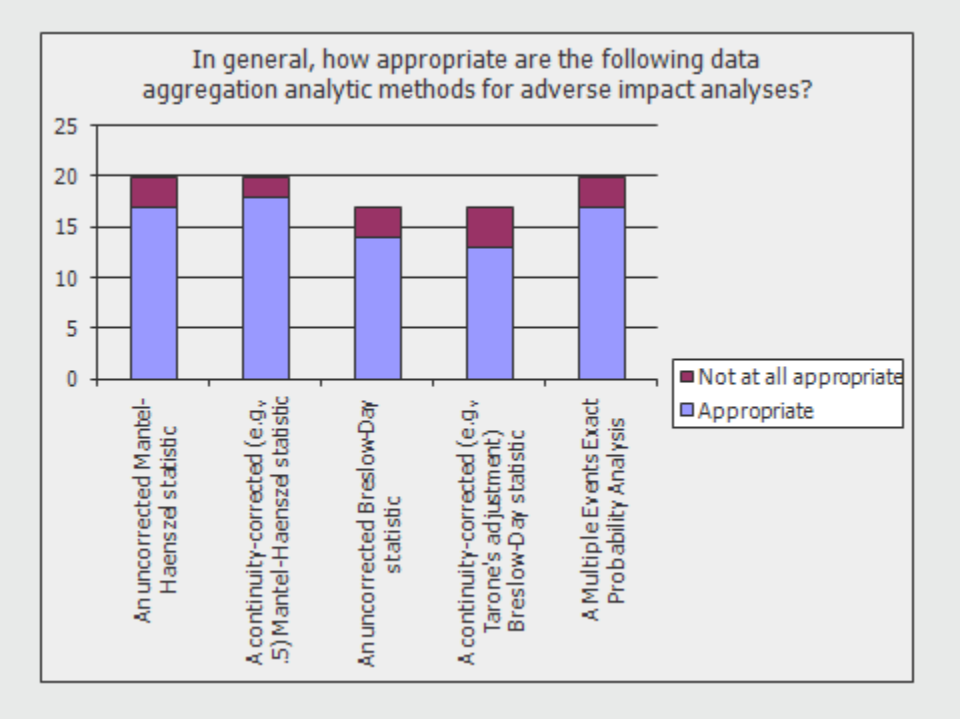
Figure 3.7 presents survey results of the frequency in which TAC members use various data aggregation statistics, whereas Figure 3.8 presents member perceptions of the appropriateness of those statistics. As results show, the majority of respondents who answered this question endorsed using various forms of "Breslow-Day type" tests that empirically assess whether the disparity was similar across strata, and various "Mantel-Haenszel type" tests that appropriately weight disparity by strata, thus avoiding the possibility of a Simpson's paradox effect. Overall, 31 TAC members answered the frequency of data aggregation method survey question, while 23 TAC members answered the appropriateness survey question. Note that there was little variability in perceptions of the appropriateness of these methods. Whether a continuity-corrected estimator test, an uncorrected estimator test, or a multiple events exact test is used, these methods were endorsed as potentially appropriate by TAC members, although context matters. Note that this is similar to the earlier TAC decision not to endorse particular statistical significance tests in the 2-by-2 case. In the focus group, the appropriateness of the statistical model (e.g., hypergeometric vs. some form of binomial) was once again discussed, and the consensus was that the decision regarding the most appropriate aggregation test should be based on how the employment decisions were made.

Figure 3.7: Frequency of various aggregation statistic use



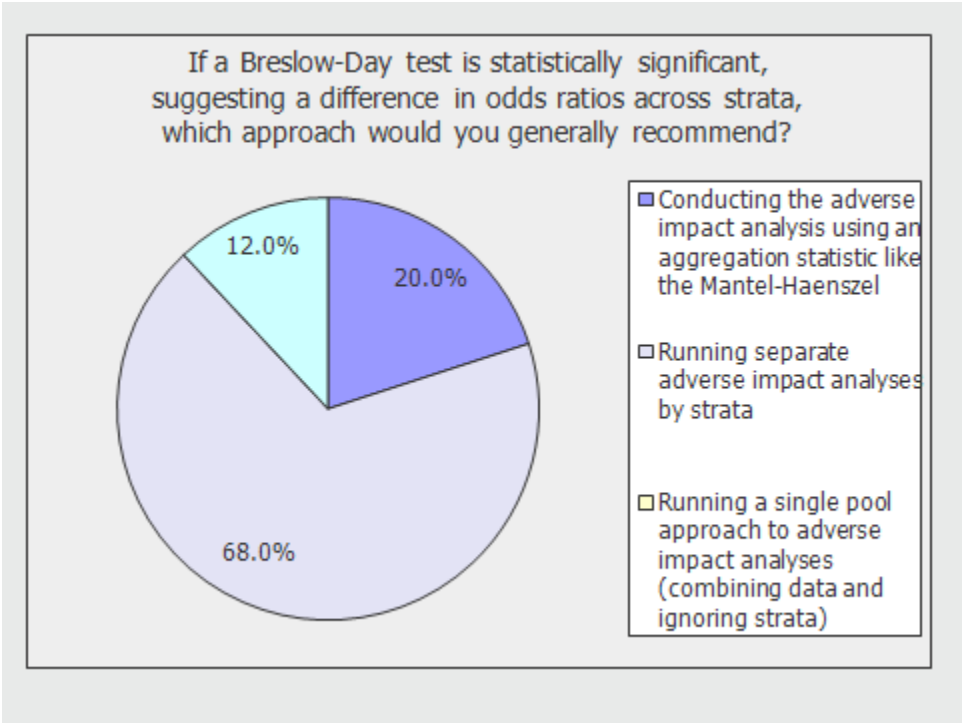
Focus group members did agree that aggregation statistics were critical, and shared examples where combining data across meaningful strata either masked a meaningful disparity or inflated a trivial disparity. Participants also emphasized the importance of an "eyeball test" (e.g., a descriptive review of section rates and effect sizes across strata) in these situations, particularly where sample sizes vary substantially and a Breslow-Day test may be less useful because of its dependence on sample size. However, an eyeball test does not account for sampling error. Some participants did mention that they use a Mantel-Haenszel approach but not a Breslow-Day analysis, under the rationale that the Mantel-Haenszel test is valid regardless of Breslow-Day results.

Figure 3.8: Appropriateness of various aggregation statistics



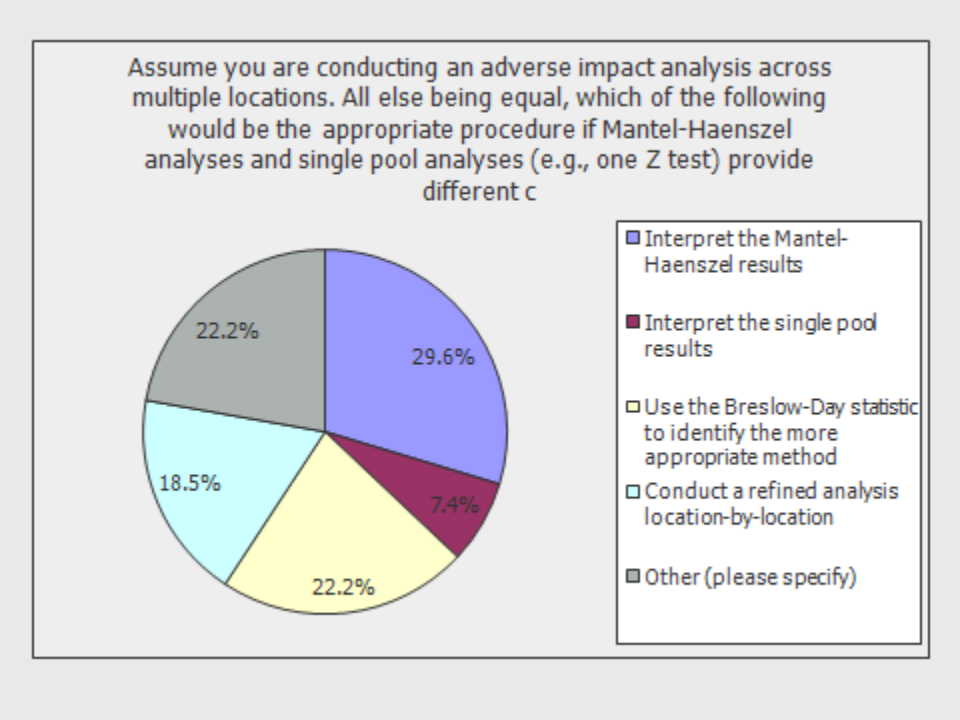
One specific question asked of TAC members was the following: If a Breslow-Day test is statistically significant, suggesting a difference in odds ratios across strata, which approach would you generally recommend? Figure 3.9 presents results of the 25 TAC members who answered this survey question. The vast majority of respondents (68%) advocated running separate analyses across strata, while 20% endorsed conducting the analysis using a "Mantel-Haenszel type" analysis. Note that no respondent advocated the use of a single pool approach in this context. Focus group participants agreed with survey results, and suggested using a descriptive approach to understanding differences in disparity across strata and why those differences may exist, and modeling strata appropriate after these considerations.

Figure 3.9: Approaches when a Breslow-Day type statistic is statistically significant



One other specific question TAC members were asked was the following: If Mantel-Haenszel type results and single pool results (e.g., a binomial Z-test) provide different conclusions (e.g., one is statistically significant and the other is not), how should results be interpreted? Figure 3.10 presents these results, and 27 TAC members answered this question. Results were split such that about one-third of respondents to this question would interpret Mantel-Haenszel type results, about a quarter would use the Breslow-Day to point them toward the more appropriate results, and about one-fifth of respondents would conduct the analysis strata by strata, and 7% endorsed a single pool approach.

Figure 3.10: Interpreting results when Mantel-Haenszel type and single pool results are inconsistent



Recommendation 3.9: *The TAC decided to make a general set of recommendations regarding issues of data aggregation.*

- *Both theoretical and empirical considerations should be made in determining whether aggregation is appropriate;*
- *Data aggregation statistics should be used when analyses are conducted across strata. This includes statistics that are used to determine whether (1) disparity is similar across strata (and thus whether aggregation is reasonable) and, if aggregation is reasonable, (2) that weight disparity appropriately across strata;*
- *In some situations a single pool approach may mask or inflate disparities and should not be used. In these situations a Mantel-Haenszel type approach or separate analyses by strata should be used instead; and*
- *When a Breslow-Day type test is statistically significant, some method to account for differences across strata is recommended. A common approach is to conduct*

separate analyses for each strata, assuming sample sizes are sufficient to permit such analysis.

One last question that was added to the agenda by focus group participants in this section concerns dependence of observations. Specifically, multiple participants wanted to discuss how multiple and frequent applicants should be treated in analyses. In this situation, knowledge of how an applicant fared for one employment decision likely affects the likelihood of another selection event for that applicant, and as such multiple applications from the same person are dependent. This dependence violates various assumptions of common statistical significance tests, and in such cases, frequent applicants may leverage results in unexpected ways.

A number of strategies for handling this situation were discussed, including the development of identification criteria for frequent applicants, data management considerations (e.g., include frequent applicants in some but not all pools), and using logistic regression approaches (e.g., via generalized estimating equations) that produce more appropriate standard errors. Participants agreed that removing duplicate observations is reasonable, but that separate applications from the same person should usually be controlled statistically, because this was the reality of the hiring process. Although the group could not come to agreement on what to do, participants agreed that considering dependence of observation was critical and that more research in the scholarly literature should focus on this issue. Participants shared stories where a frequent applicant substantially leveraged analyses in various ways, either masking or inflating results.

More Complex Analyses

One issue that was mentioned numerous times during the focus group concerned the notion that the assumptions of simpler statistical models are often not met. In many situations where unstructured employment decision processes are used, subjective human judgment is used to make decisions. For example, a candidate's education or experience

could be used to make employment decisions. In simpler binomial and hypergeometric models, the only variable that is linked to the likelihood of an employment decision is protected group. That is to say, the research question is bivariate in nature, and focuses on whether gender, race, or other protected group status "predicts" the likelihood of an employment decision. However, when other information related to the likelihood of selection is available, it may be reasonable to control for this information in statistical analyses, and assess whether protected group status predicts the likelihood of selection after accounting for legitimate and non-discriminatory factors.

In the statistical literature, this situation is called the "omitted variable" problem or "model misspecification" error, and essentially means that there are other variables that should be included in statistical analysis predicting an employment decision. Note that in the EEO context, this slightly changes the question being analyzed. In the simple 2-by-2 case, the question of interest is whether protected group and the likelihood of an employment decision are correlated. When other variables are considered, the question becomes whether protected group and the likelihood of an employment decision are correlated after already accounting for legitimate and non-discriminatory factors.[13]

The appropriateness of controlling for these factors was discussed during the focus group. In general, participants suggested that controlling for these factors was reasonable if (1) the "bottom line" employment decision process was unstructured and did not include separate and identifiable steps/components (e.g., a test, interview, work simulation), and (2) if the explanatory variables (e.g., education, experience, whether the candidate was

[13] Note that there are two general ways to control for other factors. One way is to use that other factor as a strata variable and conduct and/or aggregate analyses separately. A second option is to model other variables as part of a regression equation. One of the advantages of using a regression approach is that the analysis includes information on whether that other factor was an explaining factor in the likelihood of an employment decision. Note that, as always, the goal of mirroring reality should drive analyses, and that some factors (e.g., job, location, year) may be more appropriately controlled via aggregation analyses, while other factors (e.g., education, experience) may intuitively make more sense as a predictor in a regression equation.

internal or external) were not themselves "tainted" by potential discrimination.[14]

When no specific practice/policy is examined, controls make sense to compare only similarly situated (equally likely to be selected) individuals via measures of qualification. Some participants noted that, if a structured step in the decision process (e.g., physical ability test, interview) is identified as the cause of adverse impact in a traditional disparate impact case, it is nonsensical to statistically control for other explanatory factors because we know that the step in the process was the cause of the disparity. In legal terms, it appears that controlling for explanatory variables in a pattern or practice scenario where unstructured decisions are made is more appropriate, whereas controlling for factors when an actual assessment is used to make decisions is less appropriate, because this scenario represents an adverse impact scenario. In other words, if impact exists from results of an assessment, the assessment must be justified via business necessity or job relatedness. It is not a probative question as to whether other factors explain the disparity, because the assessment has already been shown to be the adverse impact "trigger", and the assessment user is burdened with demonstrating job-relatedness/business necessity. In some narrow situations, statistics might help to focus the validation research (e.g., if a structured interview is intended to measure education and experience and produces meaningful impact, measures of those factors may explain the disparity but the interview must be shown to be job-related).

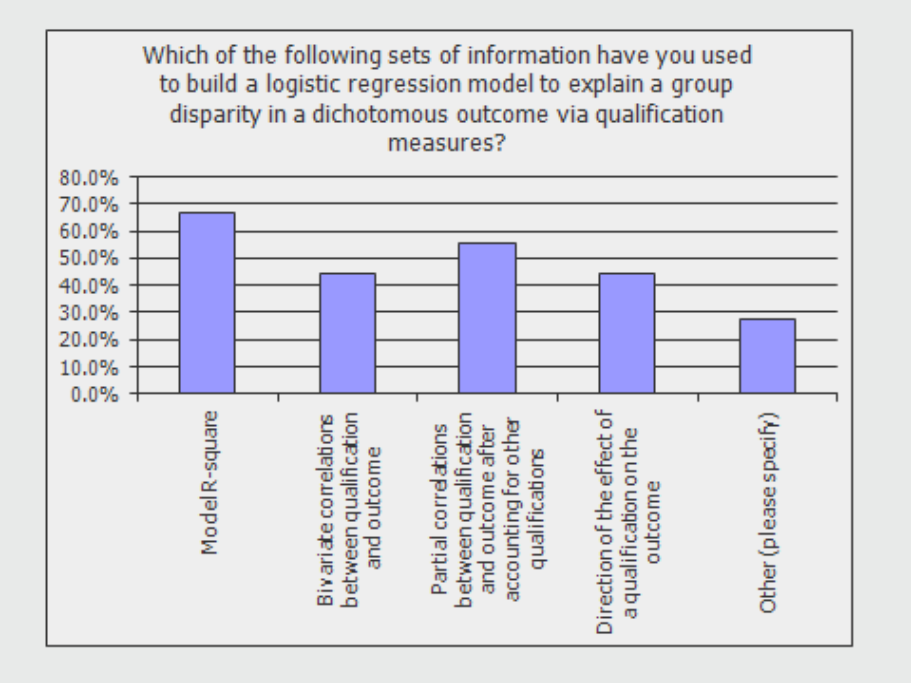
TAC members were also asked whether they use logistic regression analyses to control for other factors in a disparity analysis. Sixty-seven percent of the 36 respondents had used logistic regression. Figure 3.11 shows the different methods used to build models with explanatory variables. As results show, a variety of model building strategies were used by the 18 TAC members who answered this question, including those that focused

[14] For example, when there is an allegation of discrimination in promotion decisions in favor of men, a performance appraisal tool may be in part used to make those decisions. These ratings could be used to explain differences in promotions. However, if the performance appraisal system is unstructured or unmonitored, discrimination (i.e., male supervisors intentionally rating females lower than males on account of their gender) could be used in making the actual performance appraisal ratings. In this situation the performance appraisal tool may be tainted by discrimination, and cannot be used to explain employment decisions.

on the predictive adequacy of the overall model, as well as the effects of each variable separately (both in a bivariate context and after accounting for other variables in the model). Focus group discussion concluded that there are no concrete rules of thumb for model building, and that both conceptual and statistical considerations should be made where attempting to mirror reality.

Interestingly, the majority of focus group participants suggested that this analysis is probative in pattern or practice cases, and that similar analyses should be conducted more often than they are. This led to a detailed conversation of the assumptions of simpler models, how those assumptions are violated in the vast majority of cases, and how logistic regression models could be meaningful, and, perhaps most importantly, answer a more appropriate question related to potential discrimination. Participants also noted that this rationale of modeling other explanatory factors is well accepted in the compensation discrimination context. However, participants noted that (1) building a database with reliable information can be very time consuming and expensive in this context, (2) models may be meaningless if care hasn't been taken to mirror reality and code data reliably, and (3) it may be challenging to explain logistic regression results to decision makers.

Figure 3.11: Factors used to build logistic regression models



Recommendation 3.10: *Logistic regression analyses that model explanatory factors are very useful when subjective employment decision systems are used. These analyses should be conducted more frequently and may be useful tools available in a variety of contexts. In fact, such analyses may provide meaningful insight into perhaps a more probative question: is protected group status related to the likelihood of selection after accounting for legitimate and non-discriminatory factors? However, care should be taken to (1) build appropriate databases, (2) model reality, (3) ensure that the factors themselves are legitimate and non-discriminatory, and (4) control for capitalization on chance.*

One final question from this section asked focus group members to list the most common flaws that they saw in adverse impact analyses. Overall, 26 TAC members answered this open-ended survey question. Results are found below:

- Over-aggregation and over-generalization of trivial statistical analyses when sample sizes are very large;

- Focus on only practical or statistical significance:
    - Too much reliance on the 4/5th rule or on simple 2-by-2 tables can lead to inaccurate conclusions;

- Interpreting non-significant results in small samples as indicating no adverse impact, rather than lack of certainty;

- Interpreting significant results in large samples as meaningful adverse impact, ignoring practical significance;

- Data errors - inaccurate data/counting the wrong numbers: improperly forming the applicant pools and improperly determining the selections:
    - Failure to model the actual selection process;

- Strategically aggregating or isolating protected groups:
    - Comparing two racial subgroups (for example, Black vs. Whites) in isolation, without somehow taking into account the other individuals who do not belong to these two subgroups. There are appropriate statistical methods for examining, say, Blacks vs. Whites, while including, rather than excluding, individuals in other subgroups;
    - Combining multiple racial/ethnic groups into various aggregates;

- Mathematical errors and calculations are common errors. Not so much using the wrong test but more data entry, misreading tables, or using the wrong formulas; and

- Inclusion of irrelevant variables in the regression analyses.

Summary

This section of the TAC obviously covered many important issues, and all of these issues may drastically affect what statistical methods are used, what other factors should be included in analyses, how analyses should be aggregated, and whether statistical analyses model the reality of employment decision-making. At the end of the focus group participants were asked what themes they thought were most important in this section. The following four themes emerged:

1. Multiple adverse impact measures should be used in many situations. The TAC generally accepts statistical significance tests, and sees value in practical measures (particularly effect sizes capturing the magnitude of disparity). The $4/5^{th}$ decision rule is not well accepted by the group because of poor psychometric properties, and may only be computed today because UGESP still exist.

2. The statistical model that best represents the reality of the employment decision system should be used. This may be a hypergeometric model or some form or

binomial. However, in many situations it is difficult to clearly understand which model best fits the reality of employment decision making. The TAC recommends that analysts understand the reality of employment decision making, and choose the appropriate model or test from there.

3. Aggregation is a serious issue, and statistical and theoretical considerations should be taken into account when deciding on level of aggregation. In situations where there are strata of interest, statistical methods should be used to determine whether data should be aggregated, and to appropriately weight strata level data if appropriate. In many cases the single pool approach may be misleading. The TAC endorses the use of multiple event methods.

4. In many situations simpler 2-by-2 table analyses ignore important information about legitimate explanatory factors. In some situations logistic regression analysis may be the more appropriate analysis to assess the potential for discrimination. The TAC endorses the use of logistic regression in this context.

IV: Legal and Policy Issues related to Adverse Impact Analyses

Determining whether a disparity in selection rates is meaningful is sometimes more of an art than a science. Typically when you ask legal counsel whether the analysis results equate to a finding of discrimination you will get an answer of "it depends" and that "context always matters." This response may be due in part to the desire of the legal community to reserve the right to conduct the analyses in "different ways" to determine which way puts their client in the best light. However, all of the attorneys in the TAC agreed that a best practices document was in the best interest of the EEO community and consistency and guidance is very important. In this section, we will tackle multiple issues that the legal community deals with when trying to interpret results or guide a client through the process of conducting an adverse impact analysis.

The section will deal with the following questions:

1. What constitutes a selection decision under the *Uniform Guidelines on Employee Selection Procedures*?
2. Which types of employees should be included in the adverse impact analyses?
3. How many unsuccessful attempts to contact the job seeker does an employer have to make before the employer can treat him/her as a withdrawal?
4. Should both internal and external applicants be included in the same selection analysis?
5. What constitutes a promotion?
6. Is the statistical analysis for a disparate treatment pattern or practice case different than for a disparate impact case?
7. How do we determine what is actionable adverse impact?
8. What is practical significance?
9. When is it appropriate to aggregate data?
10. What is a pattern of discrimination?
11. What are appropriate methods for calculating a shortfall?

What Constitutes a Selection Decision under the Uniform Guidelines on Employee Selection Procedures?

Determining what constitutes a selection decision under the UGESP is a commonly misunderstood issue. Most employers believe that only results stemming from a standardized test (e.g., paper-and-pencil exam) are subject to adverse impact and validity standards. However, a close reading of the UGESP makes it clear that anything used to make a selection decision (e.g., pass/fail, select/not-select) is subject to adverse impact and validity standards. More specifically, the UGESP state:

> *"A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact."*

A majority of the focus group members agreed that anything used to make a selection decision was subject to UGESP standards. Some members expressed concern that although subjective decision making (e.g., a resume review or interview without a clear understanding of what worker characteristics are important to the job and no structured scoring system) was covered under UGESP, it is virtually impossible to validate unstructured and subjective processes.

The focus group also raised the issue of whether the UGESP, in addition to selection procedures, also cover recruitment activities. The group reached consensus that the UGESP made it clear that procedures used to "stock" a pool are considered recruitment strategies and therefore are not covered. For example, such activities as career fairs, job fairs, and online postings would not be covered. A good rule of thumb is that anything used to stock the pool is not covered, whereas anything used to reduce the pool is covered as a selection procedure.

Finally, there was a small minority that believed that recruitment could be covered under a disparate treatment pattern or practice versus a disparate impact theory. The minority believed that much has changed since the release of the UGESP and some members thought it was feasible to apply an impact theory to a recruiter who used a search strategy (e.g., use of zip codes) that disproportionately screened out members of a protected class.

Recommendation 4.1: *Anything used to make an employment decision may be considered a selection procedure and should be monitored for adverse impact. If impact is identified, the selection process should be validated in accordance with the UGESP.*

Type of Employees to be Included in the Adverse Impact Analyses

When conducting adverse impact analyses, there can be confusion regarding whether certain classes of employees (e.g., temporary employees, interns) should be included in the analyses. To properly make this determination, it is necessary to examine the context in which the matter is being evaluated. That is to say, the EEOC[15] and OFCCP[16] have different definitions of an "employee". The term employee was originally defined under the National Labor Relations Act (NLRA), which served as the source of the definition under Title VII. If a person is employed 90 days or less, he/she is not considered an

[15] "Employee" refers to any individual on the payroll of an employer who is an employee for purposes of the employers withholding of Social Security taxes except insurance sales agents who are considered to be employees for such purposes solely because of the provisions of 26 USC 3121 (d) (3) (B) (the Internal Revenue Code). Leased employees are included in this definition. Leased Employee means a permanent employee provided by an employment agency for a fee to an outside company for which the employment agency handles all personnel tasks including payroll, staffing, benefit payments and compliance reporting. The employment agency shall, therefore, include leased employees in its EEO-1 report. The term employee SHALL NOT include persons who are hired on a casual basis for a specified time, or for the duration of a specified job (for example, persons at a construction site whose employment relationship is expected to terminate with the end of the employees' work at the site); persons temporarily employed in any industry other than construction, such as temporary office workers, mariners, stevedores, lumber yard workers, etc., who are hired through a hiring hall or other referral arrangement, through an employee contractor or agent, or by some individual hiring arrangement, or persons **(EXCEPT** leased employees) on the payroll of an employment agency who are referred by such agency for work to be performed on the premises of another employer under that employers direction and control.

[16] A person employed by a Federal contractor, subcontractor or Federally assisted construction contractor or subcontractor.

employee (according to the NLRA). However, to determine employee status, one must also examine other statutes (e.g., Railway Labor Act) that are potentially relevant to a given industry.

In addition to the EEOC and OFCCP definitions of an employee, one must refer to Title VII and the courts when it comes to the issue of joint employment. For example, imagine a company that has a formal contract with an outside recruiting firm that requires the recruiting firm to source and select candidates for the company to review. The courts and Title VII have been clear that the company may have both joint employer obligations as well as non-delegable duty contracts[17] with firms that are acting as its agent. Courts have recognized joint employers when separate business entities "share or co-determine those matters governing the essential terms and conditions of employment." Whether an entity has sufficient control over workers to be deemed a joint employer turns on factors including the authority to hire, fire, and discipline employees; to issue work assignments and instruction; and to supervise their daily activities. In addition, the non-delegable duty under the law says that you are not at liberty to delegate certain duties such as hiring and recruitment. This is found in corporate labor law.

Table 4.1: Survey results about which types of workers should be included in adverse impact analyses

Type of "worker'	Include	Do not include	Response Count
Temporary workers hired from an outside firm	14.00%	86.00%	50
Temporary workers on the employer's payroll	47.80%	52.20%	46
Interns	34.80%	65.20%	46
Employees working outside the United States	22.20%	77.80%	45
Contract employees (i.e., under a 1099 provision)	19.60%	80.40%	46

[17] That is to say, organizations cannot delegate EEO responsibilities and/or liability off onto the recruiting firm.

Other issues that were agreed upon in this portion of the focus group included the following:

- A 1099 is a contract issue and not an employment issue;
- The decision maker (i.e., organization) is responsible for the employment decision; and
- The organization isn't responsible for the applicant flow data of temporary employees in some situations. An applicant to a temporary agency only becomes an applicant to an organization when it considers the candidate for employment because it is only then that the organization is making a selection decision.

The focus group reached consensus on two other related issues. (1) Much of who should be included in this context depends on the situation and also depends on the statute being covered. (2) Either party (whether a temporary service or a contractor) must take ownership of its applicant flow data. That is, it is inappropriate to put the responsibility onto another party.

Recommendation 4.2: *An employer needs to make a decision in terms of which context they are conducting adverse impact analyses, and that decision make affect analyses. For example, in the context of affirmative action plans most TAC members agreed that temporary workers not on a company payroll, interns, contract employees and those individuals working outside of the United States would not be included in the affirmative action plan and corresponding adverse impact analyses. However, under different statutes and regulations it may be appropriate to include those individuals. Finally, it is a best practice to monitor the selection practices of those organizations that are acting on behalf of or as an agent of its company.*

Table 4.2: Example of different strategies for analyzing internal and external applicants

Strategy 1: Combine Both Pools					
	Pool	**Selected**	**% Selection**	**80% Test**	**Statistical Significance**
Men	**100**	**20**	**20%**	**30%**	**2.94**
Women	**100**	**6**	**6%**		
Strategy 2: Separate Analyses					
Internal Pool					
	Pool	**Selected**	**% Selection**	**80% Test**	**Statistical Significance**
Men	**40**	**0**	**0%**	**N/A**	**N/A**
Women	**70**	**0**	**0%**		
External Pool					
	Pool	**Selected**	**% Selection**	**80% Test**	**Statistical Significance**
Men	**60**	**20**	**33%**	**60%**	**1.31**
Women	**30**	**6**	**20%**		

How many unsuccessful attempts to contact the job seeker does an employer have to make before the employer can treat him/her as a withdrawal

Those individuals who remove themselves from the selection process may reasonably be removed from the adverse impact analysis. There are instances when a company identifies a qualified candidate in which it is interested in pursuing but may have difficulty in contacting that candidate for further consideration (e.g., phone screen, interview). How many attempts does an employer have to make to contact a job seeker before the employer can consider them a "withdrawal" for purposes of the analysis?

Approximately 56% of all TAC survey members said that individuals can be removed from the analysis after one attempt and that number jumped to over 90% when job seekers were contacted multiple times. Most focus group members who were in favor of removing individuals after one attempt cautioned that this is only true if it was the policy of the company to make only one attempt and if this policy was consistently applied to all job seekers. For example, an organization would not be justified in removing one applicant from the analysis after one attempt while it hired another applicant after multiple attempts.

Although there was no clear consensus on the issue of removing someone after one attempt, most focus group members felt more comfortable with the removal after one attempt if the practice was consistently applied. It should be noted that although there is no legal standard across agencies, the position of the OFCCP, for example, is that at least two attempts must be made before an employer can presume withdrawal.

Recommendation 4.3: *Companies should establish a policy for how many attempts it needs to make before a job seeker can be deemed a withdrawal. Although no minimum threshold was established by the TAC, most members agreed that a company could justify removing someone after one attempt as long as it is consistently applied.*

Internal and External Applicants

An issue commonly faced by employers is whether to co-mingle internal employees and external applicants in the same selection analysis when both are competing for the same position. This practice has become more common with the advent of Applicant Tracking Systems (ATS) that make it easy for a company to open a requisition and allow both internal and external applicants to apply for the same position. For such openings, companies typically use one of two strategies:

- Post internally first, evaluate the internal candidate pool, and if no viable candidates are found, open the requisition externally; or
- Post and allow both internal and external applicants to apply at the same time.

When conducting adverse impact analyses of such positions, an organization must decide whether to include both internal and external candidates in a single analysis or to conduct separate analyses for both internal and external applicants. The logic behind conducting separate analyses is that in theory, the analysis of internal applicants is a promotion analysis of similarly situated employees whereas the analysis of external applicants is an external selection analysis of similarly situated applicants.

To illustrate the dilemma, consider the following example where a company posts a requisition and allows both internal and external candidates to apply. Assume that there are 110 internal candidates (40 males and 70 females) and 90 external job seekers (60 males and 30 males). Of those 110 internal applicants, none were selected and from the 90 external job seekers, 26 (20 males and six females) were selected. In addition, consider the following facts:

- The requisition is opened and both internal and external candidates apply for an open position; and
- Both internal and external go through the same selection process; and

- Internal candidates fill out a scaled down application form specifically for internal candidates.

As mentioned previously, there are two strategies to analyze the applicant flow data for this data set. Strategy 1 is to combine both internal and external applicants in a single pool approach and conduct the analysis. Strategy 2 is to separate the two pools and conduct a promotion analysis separate from an external selection analysis.

Which analysis is most appropriate? In situations where internal applicants are only competing against other internal applicants, 92% of survey respondents indicated that they would not include them in an external applicant analysis. Instead, TAC members recommended a promotion analysis on these data. However, when both internal and external applicants are competing against each other for a position, 92% of TAC survey members indicated that they should be grouped together for a selection analysis.

Focus group members further elaborated on this and said that the UGESP speak to a "selection" analysis and do not mention anything about a "hiring" analysis. One focus group member further elaborated on this issue and stated "whether or not they are running in the race is all that matters". Meaning, if they are all competing for the same job and they are being evaluated they should be analyzed together.

A small minority expressed concern with this strategy and felt strongly that internal and external applicants are not similarly situated and should never be included in the same adverse impact analysis regardless of the practice of simultaneously evaluating both internal and external applicants at the same time.

Recommendation 4.4: *When internal and external job seekers are apply together for the same requisition it is reasonable to analyze them together to evaluate the impact of the selection process. However, if they are not being considered simultaneously it is reasonable to keep them as two different pools for analysis.*

What constitutes a promotion?

When an organization pulls data from its system, it typically must set parameters that define what constitutes a promotion. In setting these parameters, the organization must first consider the type of promotion of interest. There are at least two types of promotions that could be analyzed either separately or together. The first type is commonly referred to as a natural progression promotion. In this instance, an internal employee is considered to be "in line" for a promotion due to a natural progression of the job. As such, there is no posted requisition and the employee does not formally apply for the promotion. The other type of promotion is a competitive promotion. Unlike a natural progression promotion, the opening is formally posted, and an internal employee specifically applies for that open position, usually through the organization's Applicant Tracking System (ATS).

Posting policies will differ from one company to another and understanding the policies and procedures is usually helpful when defining a promotion analysis. However, for non-competitive promotions, there was clear consensus on the definition of a promotion: 80% of respondents said that a promotion is an increase in responsibility with a job title change and an increase in pay. This is consistent with EEOC's definition of a promotion which is a change in content, scope, and responsibility.

It should be noted that during the focus group session, the group discussed the method in which an analysis of promotions is conducted. Most federal contractors and sub-contractors conduct annual promotions analyses in accordance with the requirements of 60-2.17. The typical promotions analysis consists of evaluating the number of promotions that occurred from and within a job group during the affirmative action plan year. This total number of promotions, typically both competitive and non-competitive, is then divided by the number of incumbents in the job group at the beginning of the year.

Consider the following example:

Total # of Promotions = 10
Total # of Female Promotions = 3
Total # of Male Promotions = 7
Total # of Males in the Beginning of the AA Plan Year = 100
Total # of Females in the Beginning of the AA Plan Year = 100

Analysis of Promotions
Male = 7/100 = 7%
Female = 3/100 = 3%

Most of the focus group members agreed that this isn't necessarily the most appropriate method for comparing promotion rates. First, this analysis assumes that all of the employees in the denominator are eligible, interested, and qualified for a promotion. In addition, this method of calculation does not differentiate between a competitive and non-competitive movement.

The recommendation from the group for non-competitive promotions was to conduct some sort of timing or survival analysis that would more accurately analyze non-competitive promotions. That is, are there gender or race/ethnicity differences in how long it takes to be promoted?

Recommendation 4.5: *For purposes of conducting adverse impact analyses, a promotion can reasonably be defined as an increase in responsibility with a job title change and an increase in pay. In addition, TAC members agreed that it does not make sense to include both natural progression promotions with competitive promotions in the same analysis. Finally, it would be more appropriate to analyze natural progression promotions using some sort of timing and/or survival analysis versus a typical 2-by-2 adverse impact analysis.*

Disparate Treatment[18] Pattern or Practice versus Disparate Impact

The question presented to the group was whether the statistical analysis for a disparate treatment pattern or practice case is different than for a disparate impact case. The survey results showed that 60% of respondents said that from a statistical perspective, the two types of analyses are different. However, the focus group reached consensus that the actual statistical analysis is the same but that the material facts of the case and the ultimate burden of proof were very different.

It is important to understand the difference between the two legal theories and how the burden of proof changes with each theory. Disparate impact theory operates under the notion that a facially neutral practice that is applied consistently to all groups produces a statistical disparity against a particular group. In this instance, the burden shifts to the employer to prove that the selection criteria being used are job related and consistent with a business necessity in accordance with the UGESP. If the employer meets this burden, the burden then shifts to the plaintiff to show that either the selection criteria is simply a pretext for discrimination or that there was a reasonable alternative that was equally as valid but would result in less adverse impact.

In a disparate treatment pattern or practice case, there really isn't a single neutral selection practice that is being challenged but rather, a claim that the organization's selection criteria are not being uniformly applied or that one group is being treated differently than another. To support a claim of pattern or practice disparate treatment discrimination, the plaintiff must show that discrimination, whether racial or gender, is the organization's "***standard operating procedure…that is the regular rather than the unusual practice***." (*International Brotherhood of Teamsters vs. United States*, 1977*)* Typically these cases involve some sort of statistical analysis with strong anecdotal evidence that brings that statistics to "life". Regardless, statistical analyses of adverse

[18] One other issue that came up in focus group discussion was whether the Supreme Court ruling in *Ricci vs. DeStefano* (2009) had implications for adverse impact. The focus group agreed that Ricci wasn't an adverse impact or affirmative action case, and had no direct implications on adverse impact analysis.

impact do not differ across these theories but the burden of proof and the extent of the statistical differences in selection rates do.

Recommendation 4.6: *Although the statistical methodologies used for a disparate impact and disparate treatment pattern or practice case may be the same (i.e., analyses of 2-by-2 tables), the material facts of the case and the ultimate burden of proof are very different.*

Determining what is actionable adverse impact

Determining if the results of a selection disparity analysis constitute a legal disparity under Title VII, ADEA or Executive Order 11246 is a critical issue. In other words, how much statistical evidence is enough for the results to require an employer to justify their decision-making process? As discussed in the statistics section, there are several ways in which an analyst can evaluate the differences in selection rates. These methods may include the 80% rule, statistical significance tests, and other practical significance tests (i.e., the size of the difference).

The focus group discussed this issue and could not reach consensus on how much evidence is enough. However, participants did agree that context matters. There did, however, seem to be consensus from the group that from a proactive basis, an employer should run all three tests to determine potential liability. From a legal perspective, however, the group felt strongly that an 80% violation without statistical significance was most likely not actionable adverse impact without gross evidence of intentional discrimination.

Recommendation 4.7: *Actionable adverse impact is very difficult to define in the abstract. TAC members felt strongly that context has to be taken in to account before one can feel confident that the observed differences in selection rates are actionable under the law.*

Data aggregation issues[19]

One recent trend in adverse impact litigation is the aggregation of selection procedure results across multiple locations. That is, if an organization uses the same selection procedure across multiple locations, is it appropriate to aggregate the data? As shown in Table 4.3, TAC members believe that the answer depends on the type of selection process being used: Depending on the circumstances, it can be appropriate to aggregate across locations when the selection process is highly structured. When the selection process is unstructured, a slight majority of TAC members who responded to the survey believe it is never appropriate to aggregate across locations. Note that there were many survey respondents and members of the statistical issues focus group who did not share this opinion.

Table 4.3: Survey results about the appropriateness of aggregating data across different selection procedures

Type of Selection Process	Never appropriate	Can be appropriate depending on the circumstances	Response Count
Standardized Paper and Pencil Tests	2.20%	97.80%	45
Physical Ability Test	2.20%	97.80%	45
Unstructured Selection Processes (where there are no identifiable steps)	53.30%	46.70%	45
Interview Results (unstructured)	51.10%	48.90%	45
Interview Results (structured)	6.80%	93.20%	45

It is important to note that the focus group made it clear that context always matters when making a decision on aggregation as often there are situations in which aggregating a structured selection process may not be appropriate. For example, a company may use a standardized test and test score across multiple locations. However, the demographics of

[19] Note that this issue was also discussed in the statistical methods section.

the recruitment area may drastically differ from one location to another which may lead to misleading results once the data are aggregated.

The legal issues focus group discussed a variety of others types of aggregation such as across jobs or departments or years. In general, the focus group again noted that the context and homogeneity of the units being aggregated always matters. Aggregating units that do not use the same criteria or same selection process is not desirable.

Recommendation 4.8: *TAC members felt strongly that context always matters when making a decision on whether or not applicant data can reasonably be aggregated. Aggregating data across multiple locations may be appropriate if the selection process is standardized and consistently applied from one location to another.*

Would you consider there to be meaningful adverse impact when there is statistical impact at the total minority aggregate but not by any racial subgroup (i.e., African-American, Asian)?

Most federal contractors and subcontractors that are required to develop affirmative action plans have historically conducted their adverse impact analyses in a fashion similar to the way they conduct their utilization analysis. That is, the analysis is conducted by looking at the impact at the "total minority" level. Total minority is an aggregation of all of the historically underrepresented groups which include Black, Hispanic, Asian, Native American, Native Hawaiian, and Two or More Races. The question remains whether or not this method of aggregating subgroups presents meaningful analyses or is this just an "ease of use" aggregation method.

Fifty-six percent of survey respondents said that this was not a meaningful analysis and results could be misleading. However, during the focus group meeting, the group reached consensus that a "total minority" analysis is not actionable under the law if there is not identified impact against a particular race unless there is anecdotal evidence that all minorities were discriminated against.

More specifically, the group relied upon a literal interpretation of the UGESP that specifically discusses the analysis of the highest selected group in comparison to a lower selected group[20]. In addition, UGESP discusses the issue of subgroup representation and says that any group that does not make up at least 2% of the applicant pool should be excluded from the analysis[21].

The overall conclusion was that a "total minority" analysis may be helpful for conducting proactive analyses or when sample sizes are small. However, without a disparity against a particular race, there is most likely not actionable impact under the law. Although this was not discussed in the focus group, this notion may apply to situations where multiple racial/ethnic groups are combined into a disadvantaged or favored group, unless anecdotal evidence supported this aggregation.

Recommendation 4.9: *In most situations, TAC members felt confident that a statistically significant disparity for the "total minority" aggregate without a statistical indicator for a particular protected class (i.e. Black, White, Hispanic, Asian, etc.) was not legally actionable impact.*

---

[20] Q. What is a substantially different rate of selection?

A. The agencies have adopted a rule of thumb under which they will generally consider a selection rate for any race, sex, or ethnic group which is less than four-fifths ($4/5^{ths}$) or eighty percent (80%) of the selection rate for the group with the highest selection rate as a substantially different rate of selection. See Section 4D. This $4/5^{th}$ or 80% rule of thumb is not intended as a legal definition, but is a practical means of keeping the attention of the enforcement agencies on serious discrepancies in rates of hiring, promotion and other selection decisions.

For example, if the hiring rate for whites other than Hispanics is 60%, for American Indians 45%, for Hispanics 48%, and for Blacks 51%, and each of these groups constitutes more than 2% of the labor force in the relevant labor area (see Question 16), a comparison should be made of the selection rate for each group with that of the highest group (whites). These comparisons show the following impact ratios: American Indians 45/60 or 75%; Hispanics 48/60 or 80%; and Blacks 51/60 or 85%.
See UGESP at Q and A 11

[21] Two percentage rule

What is a pattern of discrimination?

An important issue in adverse impact analyses concerns the determination of the highest selected group in a given analysis as this will determine which group is the referent group as well as who is being "favored" in the selection process. One issue that often arises in adverse impact analyses is how to interpret results when the pattern of who is the highest selected class changes from year to year, from location to location, or from job group to job group.

Survey respondents were given the following scenario: In 2007, White applicants are the highest selected group and Black applicants are adversely impacted. In 2008, Black applicants are the highest selected group and there is adverse impact against Hispanics. In 2009, Hispanics are the highest selected group and White applicants are adversely impacted. The question is whether or not this inconsistent pattern can be used as evidence of discrimination?

Sixty-five percent of survey respondents said that this could be evidence of discrimination. Furthermore, focus group members discussed this issue in more detail and came to the conclusion that context matters and it is possible that from a disparate treatment perspective, things could have changed from year to year. For example, the hiring managers could be different across years and thus, so could the racial or gender preferences that might be tied to a given hiring manager. Likewise, there could be a shift in demographics in a certain area that may significantly impact the applicant flow data and corresponding statistical results. Again, the overall conclusion of the focus group was that the statistical results have to be interpreted in the context of the situation that might permit each year or unit to be treated as a distinct adverse result.

Recommendation 4.10: *When conducting adverse impact analyses over multiple years, it is possible to find that the statistical indicators against a protected group may shift from one year to another. TAC members discussed this issue and came to the conclusion that this could be used as evidence of discrimination in a disparate treatment pattern or*

practice case. However, members cautioned that context always matters and it is important to identify the change in organization policy, practice or procedure that would explain the different patterns.

Appropriate methods for calculating a shortfall

Calculating a "shortfall"[22] may be an important part of determining practical significance as well as the potential liability if there is a finding of discrimination. In a pattern or practice or disparate impact case the shortfall is used to determine damages for the class. Therefore, the precise calculation of the shortfall is very important. There are several different methods for calculating a shortfall. They include calculating:

- The number of hires needed to make the hiring ratios equal;
- The number of hires needed to make the impacted group's selection rate 80% of the favored group's selection rate;
- The number of hires needed to make the impacted group's selection rate 80% of the overall selection rate; and
- The number of hires needed to make the group difference not statistically significant.

Consider the following applicant flow statistics and the corresponding shortfalls based upon the shortfall theory that is applied.

[22] Note that this topic was also discussed by the statistical methods focus group. However, the majority of participants in that group viewed the shortfall as an outcome of a statistical significance test (i.e., literally the difference between the observed number of hires and the expected number of hires), and was skeptical of the shortfall as a useful measure of practical significance.

Table 4.4: Example applicant flow data and shortfalls[23]

	Selected	Applicants	% Selected	Overall Selection %
Male	30	100	30%	20%
Female	10	100	10%	

Shortfall Method	Female Shortfall
Equal Selection	10
80% of Favored Group	14
80% of Overall Rate	6
No Statistical Significance	5

As the chart below shows, different methods produce very different shortfalls. Focus group members were also split on this issue but did note that the OFCCP typically uses the equal hiring ratios strategy while most organizations want to use the shortfall to no statistical significance.

Table 4.5: Survey results concerning which shortfall method is most appropriate

Shortfall Method	Never appropriate	Response Count
The number of hires needed to make the hiring ratios equal	43.60%	17
The number of hires needed to make the group difference statistically significant	25.60%	10
The number of hires needed to make the impacted group's selection rate 80% of the favored group's selection rate	15.40%	6
The number of hires needed to make the impacted group's selection rate 80% of the overall selection rate	5.10%	2

Recommendation 4.10: *Calculating a shortfall is one of many methods used to determine "practical" significance when conducting adverse impact analyses. However, TAC members could not reach consensus on a single recommended method for calculating a shortfall.*

[23] In this example shortfalls were calculated assuming margins are fixed.

Summary

This section covered many important issues, and all of these issues may drastically affect the legal interpretations of data, statistical results, and conclusions drawn from various sources. At the end of the focus group participants were asked what themes they thought were most important in this section. The following themes emerged:

1. Anything used to make a selection decision may be considered a test and should be monitored for adverse impact. If impact is identified, the selection process should be validated in accordance with the UGESP.

2. When internal and external job seekers apply together for the same requisition it is reasonable to analyze them together to evaluate the impact of the selection process. However, if they are not being considered simultaneously it is reasonable to keep them as two different pools for analysis.

3. Although the statistical methodologies used for a disparate impact and disparate treatment pattern or practice case may be the same, the material facts of the case and the ultimate burden of proof are very different.

4. Actionable adverse impact is very difficult to define in the abstract. TAC members felt strongly that context has to be taken in to account before one can feel confident that the observed differences in selection rates are actionable under the law.

5. TAC members felt strongly that context always matters when making a decision on whether applicant data can reasonably be aggregated. Aggregating data across multiple locations may be appropriate if the selection process is standardized and consistently applied from one location to another.

6. In most situations, TAC members felt confident that a statistically significant disparity for the "total minority" aggregate without a statistical indicator for a particular protected class (e.g., Black, White, Hispanic, Asian) was not legally actionable impact in most situations.

V: General Conclusions

Many important data, statistical, and legal issues have been covered in this report, and all may drastically affect the extent to which adverse impact analyses mirror the reality of employment decisions. Determining whether selection, promotion, and termination decisions result in adverse impact is an important topic for organizations, and there is limited guidance about the specific and proper ways in which these analyses should be conducted. CCE hopes that this report fills this information gap by providing the EEO community with some technical guidance on how to most appropriately conduct adverse impact analyses. These recommendations represent a wide range of expert perspectives, including recommendations from industrial-organizational psychologists, labor economists, plaintiff's and defense attorneys, consultants, HR practitioners, and former OFCCP and EEOC officials.

One of the reoccurring themes seen throughout the document is that context matters. Employment decision making processes are often complex and care must be taken to understand those processes to ensure that data management strategies and statistical methods accurately mirror the reality of those processes. This effort often requires expertise in law, statistics, and employee selection. At the end of the day, the purpose of impact analyses is to determine whether substantial differences in employment outcomes across groups exist. Toward mirroring reality, the following are some of the more critical recommendations from this report:

- It is essential to create requisition codes for each job opening and to indicate the disposition of each job seeker;

- Not all job seekers will be considered applicants for the purpose of conducting adverse impact analyses. To be considered an applicant, a job seeker must express an interest in an open position, meet the minimum qualifications for the position, and follow the organization's rules for applying;

- Job seekers who withdraw formally (e.g., inform the organization that they are no longer interested in the position) or informally (e.g., do not return calls, fail to show for an interview) should not be considered as applicants in the adverse impact analyses;

- Applicants who are offered a position but either decline the position or do not report to work should be considered as "selections" in adverse impact analyses;

- Multiple adverse impact measures should be used in many situations. The TAC generally accepts statistical significance tests, and sees value in practical measures (particularly effect sizes capturing the magnitude of disparity). The 4/5th decision rule is not well accepted by the group because of poor psychometric properties, and may only be computed today because UGESP still exist;

- The statistical model that best represents the reality of the employment decision system should be used. This may be a hypergeometric model or some form or binomial. However, in many situations it is difficult to clearly understand which model best fits the reality of employment decision making. The TAC recommends that analysts understand the reality of employment decision making, and choose the appropriate model or test from there.

- Aggregation is a serious issue, and statistical and theoretical considerations should be taken into account when deciding on level of aggregation. In situations where there are strata of interest, statistical methods should be used to determine whether data should be aggregated, and to appropriately weight strata level data if appropriate. In many cases the single pool approach may be misleading. The TAC endorses the use of multiple event methods.

- In many situations simpler 2-by-2 table analyses ignore important information about legitimate explanatory factors. In some situations logistic regression

analysis may be the more appropriate analysis to assess the potential for discrimination. The TAC endorses the use of logistic regression in this context.

- Anything used to make a selection decision may be considered a test and should be monitored for adverse impact. If impact is identified, the selection process should be validated in accordance with the UGESP.

- When internal and external job seekers apply together for the same requisition it is reasonable to analyze them together to evaluate the impact of the selection process. However, if they are not being considered simultaneously it is reasonable to keep them as two different pools for analysis.

- Although the statistical methodologies used for a disparate impact and disparate treatment pattern or practice case may be the same, the material facts of the case and the ultimate burden of proof are very different.

- Actionable adverse impact is very difficult to define in the abstract. TAC members felt strongly that context has to be taken in to account before one can feel confident that the observed differences in selection rates are actionable under the law.

- TAC members felt strongly that context always matters when making a decision on whether applicant data can reasonably be aggregated. Aggregating data across multiple locations may be appropriate if the selection process is standardized and consistently applied from one location to another.

- In most situations, TAC members felt confident that a statistically significant disparity for the "total minority" aggregate without a statistical indicator for a particular protected class (e.g,. Black, White, Hispanic, Asian) was not legally actionable impact.

Two other points are worth noting. First, this report includes best practices at a particular point in time. All data were collected in 2009 and 2010. Law is a constantly evolving notion, and may change over time as a function of political ideology, socially derived values, court rulings, advances in science, or changes in technology. As such, the results of this TAC should be interpreted in the temporal context, and in consideration of any changes to the EEO context.

Lastly, it is important to reiterate that the information presented in this report is not a criticism of the policies or procedures of any federal agency, specific court rulings, or research from the scholarly literature. Likewise, this is not an attempt to revise the UGESP, agency compliance manuals, or any technical authority. Instead, this report includes recommendations for best practices in adverse impact analyses based on data collected from 70 experts in the field; CCE hopes that these recommendations are useful to the EEO community.

References

Biddle, D.A. (2005). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Burlington, VT: Gower.

Boardman, A. E. (1979). Another analysis of the EEOC four-fifths rule. *Management Science, 8,* 770–776.

Bobko, P. & Roth, P.L. (2010). An analysis of two methods for assessing and indexing adverse impact: A disconnect between the academic literature and some practice. In J.L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 29-49). New York: Routledge.

Roth, P. L., Bobko, P. & Switzer, F. S (2006). Modelingt he behavior of the 4/5th rule for determining adverse impact: Reasons for caution. Jounral of Applied Psychology, 91, 507-522.

Collins, M.W. & Morris, S.B. (2008). Testing for adverse impact when sample size is small. *Journal of Applied Psychology, 93,* 463-471.

Crans, G.G., Shuster, J.J. (2008). How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial. Statistics in Medicine, 27 (8), 3598-3611.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform Guidelines on Employee Selection Procedures. Federal Register, 43 (166), 38295-38309.

Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of Labor & Department of Treasury (1979). Adoption of Questions and Answers to Clarify and Provide a Common

Interpretation of the Uniform Guidelines on Employee Selection Procedures. Federal Register, 44, 11996-12009.

Gastwirth, J.L. (1988). *Statistical reasoning in law and public policy* (Vol. 1). San Diego, CA: Academic Press.

Greenberg, I. (1979). An analysis of the EEOC four-fifths rule. *Management Science, 8,* 762–769.

Gutman, A., Koppes, L. & Vadonovich, S. (2010). EEO Law and Personal Practices (3rd Edition). New York: Routledge, Taylor & Francis Group.

Lin, C.Y, Yang, M.C. (2009). Improved p-value tests for comparing two independent binomial proportions. Communications in Statistics - Simulation and Computation, 38 (1), 78 – 91.

Meier, P., Sacks, J. & Zabell, S. (1984). What happened in Hazelwood: Statistics, employment discrimination, and the 80% Rule. *American Bar Foundation Research Journal*, *1*, 139-186.

Morris, S.B. & Lobsenz, R.E. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology*, *53*, 89-111.

Siskin, B.R. & Trippi, J. (2005). Statistical issues in litigation. In F.J. Landy (Ed.), *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives* (pp. 132-166). San Francisco: Jossey-Bass.

Sobol, R., Michelson, S., Finkelstein, M., Fienburg, S., Eisen, D., Davis, F.G., et al. (1979). Statistical inferences of employment discrimination and the calculation of back pay: Part I: Inferences of discrimination. *OFCCP Statistical Standards*

Panel Report (DOL Contract No. J-9-E-8-0110, pp. 1-54). Washington, DC: U.S. Department of Labor.

Zedeck, S. (2010). Adverse impact: History and evolution. In J.L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 3-27). New York: Routledge.

Cases Cited

Connecticut vs. Teal, 457 US 440, 29 FEP 1 (1982).

Contreras vs. City of Los Angeles, 656 F.2d 1267 (9th Cir. 1981).

Frazier vs. Garrison, 980 F.2d (5th Cir. 1993).

Hazelwood School District vs. United States, 433 US 299 (1977).

International Brotherhood of Teamsters vs. United States, 431 US 324 (1977).

Moore vs. Southwestern Bell Telephone Company, 593 F.2d 607 (5th Cir. 1979).

Ricci vs. DeStefano, 129 S. Ct. 2658 (2009).

Waisome vs. Port Authority, 948 F.2d 1370, 1376 (2d Cir. 1991).