The Economic Journal, **112** (*November*), F480–F518. © Royal Economic Society 2002. Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA.

FIELD EXPERIMENTS OF DISCRIMINATION IN THE MARKET PLACE*

P. A. Riach and J. Rich

Controlled experiments, using matched pairs of bogus transactors, to test for discrimination in the marketplace have been conducted for over 30 years, and have extended across 10 countries. Significant, persistent and pervasive levels of discrimination have been found against non-whites and women in labour, housing and product markets. Rates of employment discrimination against non-whites, in excess of 25% have been measured in Australia, Europe and North America. A small number of experiments have also investigated employment discrimination against the disabled in Britain and the Netherlands, and against older applicants in the United States.

The technique of conducting carefully controlled field experiments to measure discrimination in the market place is 35 years old. Although the market is the centrepiece of the economist's attention the initial development of this technique was by British sociologists. Daniel's (1968) tests for racial discrimination in the English housing and labour market, using matched pairs of actors, were followed by Jowell and Prescott-Clarke (1970), who introduced written tests.

It was not until the 1980s that this experimental technique found a place in the economics journals, with articles by Firth (1981) in the *Industrial and Labor Relations Review*, Yinger (1986) in *American Economic Review* and Riach and Rich (1987) in *Australian Economic Papers*. Interest in field experiments of discrimination on the part of economists, did increase during the 1990s with publications appearing in several economics journals including the *American Economic Review* (Ayres and Siegelman, 1995; Kenney and Wissoker, 1994), the *Quarterly Journal of Economics* (Neumark *et al.*, 1996) and the *Review of Black Political Economy* (Bendick *et al.*, 1994). There has also been significant activity by the International Labour Office (ILO) and the Urban Institute (UI) in Washington. Consequently there is now a substantial body of literature, which demonstrates discrimination in labour, housing and product markets on the basis of sex and race. There have also been a handful of experiments investigating employment discrimination on the bases of age and disability. It is now appropriate to make the technique and its findings more widely known to economists, especially as most countries have legislated to make such discriminatory activity illegal. The experiments have been conducted by university researchers, eg, Levinson (1975), by independent research agencies – eg Political and Economic Planning, PEP; see McIntosh and Smith (1974) – and by advocacy groups – eg The Spastics Society; see Graham *et al.* (1990). In the tables detailing the results from these studies, these groups are designated respectively as U, RA and AG.

* We wish to thank the following for their assistance in compiling this survey: Frank Bovenkerk, Marc Bendick Jnr., and Elisabeth Fellowes-Smith. Further, we wish to acknowledge the helpful comments of Stephen Machin and anonymous referees on an earlier version of this paper.

Moreover social scientists have engaged in field experiments during the past three decades precisely because the alternative techniques for measuring discrimination have proved inadequate. Surveys of attitudes towards minority groups in the market are not likely to produce honest and accurate responses, as demonstrated by the classic study of La Piere (1934). Also the econometrician's application of the technique of regression analysis to published data to deduce discrimination, pioneered by Blinder (1973) and Oaxaca (1973) has been subject to considerable criticism, which revolves around the specification of the model and the choice of independent variables; see, for example, Gunderson (1989). Kim (2002) has demonstrated a difficulty with published data. She obtained conflicting results for wage discrimination against black women in the United States when she used both household and census data to calculate her estimates. A comprehensive, detailed survey and discussion of studies using this wage regression technique to test for wage discrimination on the basis of race and sex is contained in Altonji and Blank (1999).

1. Field Experiments in the Labour Market

1.1. *The Technique Explained*

Three procedures have been used to carry out direct tests for the extent of discrimination in labour markets. Two involve personal approaches; either individuals attending job interviews or applying over the telephone. The third involves responding to job vacancies with written applications. Researchers in Britain often use the term 'situation tests' to describe personal approaches whereas in the United States they are usually called 'audit tests'. Typically the term 'correspondence tests' is used to describe the technique of written approaches to advertised vacancies.

In the case of personal approaches two testers are matched; one from the majority group, say white, the other from the minority group, say black. Some tests in Britain have involved sending three matched testers (eg, a Briton, a West Indian and a Greek) for job vacancies. The qualifications and presentation style of the testers is matched as closely as possible, so that they are identical in all relevant employment characteristics and differ only in one characteristic, such as sex, race, ethnicity or disability. The matched pairs are trained in what to say in response to various questions so that both testers in the matched pair can give equivalent backgrounds to the prospective employer for such personal characteristics as schooling, qualifications and job experience. As well, two (three) brief curriculum vitae are prepared for the relevant jobs and rotated between the two (three) testers for use in their encounters with employers. Also, they may be coached in their demeanor at interview to try to control for personality differences. They are trained together so that they can closely align their manner to each other. The testers are carefully supervised throughout the testing period and in the case of telephone tests the supervisors can sit in to observe that both applicants are coming across to the employer as identical in job qualifications and experience. Aspects of the tests are recorded so that employer responses can be classified.

When the applicants go to job interviews they are given an elaborate form for recording all aspects of the interviews and these are filled in immediately on completion of the interview. The supervisor then goes through these to ensure they have been correctly filled in so that the results of a matched pair of applicants can be recorded (for example, Bendick *et al.* (1994, pp. 27–8); McIntosh and Smith (1974, pp. 7–8)). Most of the in-person tests use multiple pairs of testers to control for any unintended bias from an individual tester. The effect on the results of varying the pairs of testers as well as the effect from any one individual tester can be isolated and tested to see whether it is statistically significant.

When making the job inquiries there must be some time delay between the testers' approaches, say half an hour to an hour. It is, of course, possible that the job may be offered to another candidate in between the contacts made by the testers. Therefore the order of approach, to the employer of the testers must be considered. Many of the tests ensure that in half of the tests the first approach is made by the minority applicant and in the other half, by the majority applicant, but a number of the researchers have opted always to send the minority tester first. This ensures that, if the job is filled in between the approaches, discrimination recorded against minority applicants *cannot be overestimated*. In the British and ILO tests, the racial minority applicant always made the first approach. Any job offers are promptly and courteously declined.

It is important to distinguish three ways in which personal approaches can be used. First, they can involve direct application by telephone, and this is equivalent to written applications, which test for 'invitations to interview' only. The two applicants are identified by name and accent (or pitch in the case of women), but this is not a successful technique for testing discrimination against African–Americans (Turner *et al.*, 1991). This method has been used, for example, in Britain by Brown and Gay (1985), Hubbuck and Carter (1980), and McIntosh and Smith to test for discrimination against Indians and Pakistanis, who have distinctive names. Also Levinson used it in the United States to test for sexual discrimination. Second, personal approaches can involve direct contact with a firm, either in response to an advertisement which specified personal attendance, or as an unsolicited inquiry regarding job availability. This method has been used by Daniel (1968), McIntosh and Smith (1974), and Nunes and Seligman (1999, 2000). Third, personal approaches may involve an initial application for interview, by phone, fax or mail, followed by attendance at interview. Some minority groups, such as African–Americans are not distinguishable on the telephone, so it is only on personal contact that race is apparent. In this third variant it follows that the test is only pursued in situations where both candidates are invited to interview, at which point the employer first is confronted with racial difference. A second reason for adopting this latter technique of approach is to test all stages of the hiring process, which has been the practice of the ILO and the UI.

A problem which has been raised about personal approaches is with regard to the matching and motivation of testers. In their critique of the Urban Institute studies, Heckman and Siegelman (1993, pp. 190–1) question the effectiveness of the procedures for selecting, training and matching pairs of testers so as to ensure they are identical in all relevant employment characteristics except race: moreover,

the capacity to demonstrate tester equivalence objectively. 'This inability to defend, or even fully enunciate, the criteria used to match audit pair members constitutes the Achilles heel of the audit pair methodology' (Heckman and Siegelman, 1993, p. 191). Despite careful training of the participants in the audit pairs, it is impossible to ensure that all aspects of the applicants' performance are identical during their interaction with those performing the interview. In particular it is possible that, consciously or unconsciously, minority applicants may be motivated to prove the existence of discrimination, and thereby bias the results. The British sociologist Robin Ward first raised this criticism in 1969 soon after the first PEP study in 1967 in England (Ward, 1969, p. 220). This point has subsequently been reiterated by Heckman – 'Auditors are sometimes instructed on the ''problem of discrimination in American Society'' prior to sampling firms, so they may have been coached to find what the audit agencies wanted them to find' (Heckman, 1998, p. 104). Heckman and Siegelman conclude, quite rightly, that – 'An objective demonstration of the quality of the matches would go a long way toward making audit pairs credible' (1993, p. 271).

Heckman and Siegelman also emphasise the need to sample a range of skill levels to ensure accurate assessment of hiring behaviour (1993, p. 225), and Darity and Mason (1998) are criticised for omitting to acknowledge that hiring audits have primarily been concerned with entry level jobs in low skilled occupations (Heckman, 1998, p. 104). It is accepted by Yinger (1993, p. 269), however, that broadening the skill level will be a challenge requiring more elaborate training and role-playing.

In fact it is the case that the British developed procedures to deal with these problems in the 1960s. Political and Economic Planning's approach to selecting, matching and training testers was to employ professional actors and rely on their skills and professionalism – 'We agreed with the Equity argument that a good character actor will be better at playing an archbishop than will the Archbishop of Canterbury; that is, when it comes to playing a part, an actor is more ''real'' than a type-cast non-actor'. (Daniel, 1970, p. 354). Henry and Ginzberg (1985) also used professional actors in their Canadian tests because they too thought actors would be better able to sustain role-playing particularly when subjected to intense and lengthy interviews by employers (Henry and Ginzberg, 1985, p. 19).

The UI, however, has been critical of PEP for – 'their reliance on professional actors who were not necessarily matched to form *visually comparable* teams' (Mincy, 1993, p. 168, our emphasis). Zimmermann explains that, in the UI's hiring audits – 'Testers were matched according to objective criteria such as age, *weight, height* . . .' (Zimmermann, 1993, p. 408, our emphasis). Similarly, Fix *et al.* (1993, p. 20) assure us that UI audit pairs were – '. . .of the same *physical appearance* and personality type' (our emphasis). This raises a serious issue of, on the one hand, what variables we are controlling for and, on the other hand, what variable we are testing for. Clearly it is human capital (productivity-determining) variables which we must control for. These include education, experience, age, which are matched in curriculum vitae, and motivation, commitment and demeanor, which must be matched via careful choice and training of testers. If it is race for which we are testing we must ensure that *all* distinctive racial characteristics are signalled. Mincy

is critical of PEP for matching a single white tester against West Indian, Pakistani, Indian and Greek testers, instead of forming 'visually comparable teams'. This misses the point that your average Englishman *is* visually distinct from some of these groups in more than colour. Afro-Caribbeans, in general, are taller and heavier than the English; on the other hand Bengalis are smaller and lighter than the English. It would be an entirely artificial and unrepresentative construct to match English and Bengali testers of a comparable size. If characteristics such as height and facial hair (Heckman and Siegelman, 1993, pp. 217–8) are distinctive between the groups in the test, it is appropriate to signal this to the employer, *unless* the test is for colour, rather than race. In view of their methodological position we await with interest an UI foray into testing sexual discrimination.

Mincy's criticism of PEP for using professional actors who were not 'visually comparable' is misconceived. Moreover, it overlooks the skills which a professional actor can bring to the control of subjective components of human capital, such as, motivation, commitment and demeanor. In *Clear and Convincing Evidence* there is frequent reference to the training which testers received in the form of role-playing and mock interviews, in order to match-up their behaviour. Also Fix *et al.* (1993, p. 30) moot the future use of 'batteries of psychological and behavioural tests' to improve further this particular human capital control. Amongst the skills of professional actors is the ability to portray a variety of roles and personality types, and thereby effectively ensure this control. At the pinnacle of the profession, it has been known for an actor to portray, sequentially, a form of Manhattan low-life, a woman and an autistic.

A second British innovation was in 1969 when Jowell and Prescott-Clarke carried out written experiments. Their technique involved sending carefully-matched pairs of written job applications in response to advertised vacancies, to test for discrimination in labour hiring at the initial stage of selection for interview. In order to avoid detection, the letters obviously cannot be identical, but in all essential characteristics such as qualification and experience candidates are closely matched so that the only effective distinguishing characteristic is race, ethnicity, sex, age or disability. Moreover, to control for the possibility that the style of a particular letter might influence employer response, letter type is alternated and allocated equally between the two groups. The advantage of this correspondence technique is that the researcher is able to exercise precise control over the content of applications, to control for any unintended bias in letter type by equal allocation between the groups, and to *demonstrate* the controlled and objective nature of the procedure to the reader. This includes the publication of the standard letters of application.

Jobs to be applied for are usually chosen from daily newspapers in the geographic region. Two standard letters of application are prepared, although some of the tests have sent three standard letters to each occupation selected (Brown and Gay, 1985; Hubbuck and Carter, 1980) and one has sent twelve (Firth, 1982). The letters can be adapted to test also for the effect of some of the other control variables, such as qualifications and marital status (Jowell and Prescott-Clarke, 1970; Firth, 1982). Sex and race (with the exception of Afro-Caribbeans and African–Americans) can be identified by the applicant's name; disability by a

paragraph explaining the applicant's disability. We suggest that the practice of Adam (1981), in his sexual preference experiment, could be implemented to identify African-Americans; that is, they could be identified in the 'Interests' section of their curriculum vitae by involvement in political and cultural activities associated with Black Americans.

The matched pairs of standard application letters are posted simultaneously within two days of the advertisement appearing. To minimise inconvenience to employers, invitations to interview are promptly and courteously declined.

Heckman and Siegelman (1993, p. 229) do note two examples of written tests: Newman (1978) and McIntyre *et al.* (1980). These two studies did not, however, involve applications to advertised job vacancies. Instead *unsolicited* curriculum vitae were posted to 'potential employers'. Newman took his listing of employers from the Office of Federal Contract Compliance, whilst McIntyre *et al.* took theirs from the College Placement Manual. This procedure does not test market activity as no actual job is on offer; instead it investigates for preferential treatment in employer responses. Significantly both studies were confined to *entry-level* jobs for college graduates. In neither Heckman and Siegelman (1993), nor Heckman (1998) is there any reference to the British and Australian development of 'correspondence testing', which had been cited in an American-based journal in 1991 (Riach and Rich, 1991–2). This is a technique which tests the hiring decision, ensures strict equivalence between testers, is free of any motivational complication and enables objective documentation of the experiment. Moreover, a variety of skill levels can be incorporated in the study; the following occupations have been included in correspondence tests – gardener, accountant, computer programmer, payroll clerk, computer analyst programmer, industrial relations officer (Riach and Rich, 1987), secretary, salesperson (Daniel, 1968; McIntosh and Smith, 1974; Riach and Rich, 1991).

'Audit studies' (or 'situation tests') have also been criticised for over-stating discrimination, wherever unobserved variables have been significant. Robin Ward made this point in 1969 - 'many other factors (besides discrimination) could have helped to account for the results of the tests. While the researchers were careful to control for some of them, there is evidence that other factors were partly responsible for coloured people being turned down' (Ward, 1969, p. 220). Heckman and Siegelman likewise warn that '. . .audit studies are crucially dependent on an unstated hypothesis: that the distributions of unobserved (by the testers) productivity characteristics of majority and minority workers are identical' (1993, p. 224). 'From the audit studies, one cannot distinguish variability in unobservables from discrimination' (p. 255). We accept Heckman and Siegelman's point (p. 222) that the burden of proof in audit studies lies with their perpetrators, nevertheless we consider it is regrettable that they provide no example of what such confounding unobservables could be. The one piece of evidence which they produce is earnings functions estimated from company personnel records: despite detailed data on personal characteristics and employment history, there remains a big unexplained residual, with R-squared rarely as high as 0.6 (p. 275). The critical question is whether this residual is the outcome of productivity-determining, unobserved variables, or whether it arises from discriminatory criteria. Anyone

who has worked in a large organisation will be aware that internal promotion is frequently related to compliance, conformity and sycophancy. It is not easy to envisage how such 'attributes' could enhance productivity: instead it can be interpreted as discrimination against the independently-minded.

It would be helpful if Heckman and Siegelman could suggest what productivity-influencing, unobserved variables could explain Neumark *et al.*'s finding that women faced discrimination in high-price, but not in middle or low-price, restaurants, or Riach and Rich's finding that women faced discrimination in computer analyst programming, but not in computer programming. Written tests can (and have) control (led) for a wide range of productivity-determining criteria, and can accommodate additional relevant variables if they are forthcoming. Until Heckman and Siegelman can identify that which is relevant, but unobserved, we have a Scottish verdict; that is, 'not proven'. Robin Ward *did* see fit to identify unobserved variables '. . .restrictions on suitable tenants in private accommodation are not confined to racial differences. Sex, family position, employment and ''respectability'' are other factors which exclude large numbers of applicants' (Ward, 1969, p. 220).

1.2. *Reporting and Interpreting the Results*

It is sound academic practice to publish full details of any field experiment. This includes the procedures adopted, and complete results of all tests, broken down by occupational category where relevant. Complete results means the number of applications made, recorded by the outcome for the matched testers at each stage of the hiring process: in a study of white/black employment opportunities, this means, at the 'invitation to interview' stage recording – both rejected/both invited for interview/only the white applicant invited to interview/only the black applicant invited to interview. If the test covers the job offer stage as well, then the results for the matched testers who proceed to the interviews should be reported in the same detail for this stage - both rejected/both offered the job/only the white applicant offered the job/only the black applicant offered the job. The test outcomes should also be reported, separately, by test pairs; see also, Heckman and Siegelman (1993, pp. 197–212). If any tests are not completed the results for these should be reported separately. Such a practice of full disclosure enables any sceptical reader to appraise the quality of the data, and to apply her/his own calculations to the interpretation of the data. All records of the tests should be kept for the purposes of ensuring that the data reported can be substantiated. The findings on discrimination should, of course, be tested for statistical significance. The effect of pairs of testers or letter type on the results can be isolated and tested to see whether it is statistically significant. Many researchers have used chi-squared tests for statistical significance; see, in particular, Bovenkerk (1992). Heckman and Siegelman (1993), in an extensive discussion and examination of in-person test results in the United States, demonstrate rigorous analytical methods that provide reliable tests of statistical significance. First, tests for homogeneity across the tester pairs can establish the validity of aggregating the results for all the pairs of testers. Second, different tests may be appropriate: chi-squared tests for large

samples; an exact, small sample, binomial test; or a conditional sign test, where the conditioning event is the success of only one tester in the pair at the relevant stage of hiring (Heckman and Siegelman, 1993, pp. 197–212). Finally regression analysis can be used to analyse other variables arising in the testing which may explain discriminatory behaviour, such as firm size, location of firm, stated equal opportunity policies of firm, occupation, industry, race/sex of the interviewer, race/sex of the tester, time of tester visit.

There has been some controversy about the interpretation of the findings of these field experiments of discrimination: in particular, the incidence of discrimination detected. When two testers apply for a job, either in person or by letter, there are four possible outcomes: both offered a job or interview, only the white (male) offered a job or interview, only the black (female) offered a job or interview, neither offered a job or interview. First we must consider what it means when neither receives an offer. Does this constitute an observation of equal treatment; that is, of non-discrimination, or does it fail to provide any information about an employer's penchant for using race (sex) as a criterion in hiring decisions; that is, is it a non-observation? The initial researchers at PEP argued that it is a non-observation (McIntosh and Smith, 1974, p. 24). This interpretation was adopted by Brown and Gay (1985), the ILO (various studies), and Riach and Rich (1987; 1991). On the other hand the UI's interpretation has been that it constitutes symmetrical treatment (Cross *et al.*, 1990, p. 44). This interpretation has been adopted by the FEC (Bendick, 1996), and Neumark *et al.* (1996). The dramatic difference which this interpretation has on the calculated incidence of discrimination is demonstrated in Table 1 of Heckman's paper in the *Journal of Economic Perspectives* (Heckman, 1998, p. 105). The percentage of employment audits which recorded equal treatment, so defined, were – Chicago, 85.8: Washington, 75.1: Denver, 86.9. But more than three quarters of these observations of 'equal treatment' were where both candidates were rejected. The percentage of audits where both candidates received offers were – Chicago, 11.2: Washington, 6.6: Denver, 15.8.

We do not share the view that a rejection of both applicants represents an act of racially (or sexually) symmetrical treatment. We interpret 'discrimination' as meaning the act of giving preference to a particular race (or sex), when *consciously* confronting such racial (or sexual) characteristics in a decision-making process. There are many reasons why job applicants may be rejected before an employer has to confront race (or sex). Initial screening may be based on timing of applications, age, current employment status, etc. In particular, the state of the macroeconomic labour market will impact on the number of occasions both receive rejections. An increase in unemployment will increase the ratio of applicants to vacancies and increase the probability of rejection; thereby reducing 'discrimination'. A hypothetical example highlights the impact of these alternative interpretations of 'both rejected'.

We envisage a test for sexual discrimination conducted for two age groups. Column 2 is male-only offered; column 3 is female-only offered; column 4 is both offered; column 5 is neither-offered; column 6 is net discrimination when 'neither invited' is treated as a non-observation, that is, $(2-3)/(2+3+4)$; column 7 is net

discrimination when 'neither-invited' is treated as an observation of 'equal treatment', that is, (2−3)/(2+3+4+5).

1 AGE	2 M	3 F	4 B	5 N	6 D%	7 D%
25	40	20	20	20	25	20
50	10	5	5	80	25	5

In each case there have been 100 applications, and there is a consistent preference for men over women in the ratio of two to one. Employers are assumed to discriminate on the basis of age so this leads to 80 occasions when both 50-year-olds are rejected, compared to 20 occasions for the younger age group. If we treat 'neither invited' as a non-observation we calculate a consistent level of discrimination of 25%. On the other hand, if we treat 'neither invited' as an observation of equal treatment, discrimination falls from 20%, when we test 25-year-olds, to 5%, when we test 50-year-olds. Perhaps such a calculation could be justified by arguing that it appropriately records the outcome of sexually-symmetrical age discrimination, but if the screening acts to consign all 50-year-olds immediately to the wastebin, no conscious preference formation on a sexual basis has been involved. A similar view has been expressed by Darity and Mason (1998, p. 79). 'This is a fairly stringent test for discrimination, since, in the case where no offer was made to either party, there is no way to determine whether employers were open to the prospect of hiring a black or an Hispanic male, what the overall applicant pool looked like, or who was actually hired'.

Second, we must consider what constitutes an act of discrimination. The traditional approach of Daniel (1968), McIntosh and Smith (1974), and the ILO (Bovenkerk, 1992) is to record discrimination at the point when differential treatment first occurs, that is, when one candidate first encounters preference relative to her/his pair. This is not the position adopted by Heckman and Siegelman '...if tester A is denied an interview, while tester B is interviewed but is nevertheless rejected for the job, should one consider this outcome as evidence of discrimination?' (1993, p. 227, fn 7). This rhetorical question receives an unequivocal answer in the text – 'Given... the clear bottom-line nature of a job offer, we focus on the ''get-a-job'' measures of discrimination in this chapter'. The tables in Heckman (1998) and in Heckman and Siegelman (1993) consequently do not record as discrimination situations where one tester is rejected *for* interview, whereas the paired tester is rejected *at* interview. This practice is inconsistent with their own declaration that 'asymmetry of treatment of ''identical'' persons constitutes evidence of discrimination' (Heckman and Siegelman, 1993, p. 198). It also disregards the relative demotivational impact on any group which is consistently denied job interviews; see also, Turner *et al.* (1991, p. 32). The measurement implications are readily apparent; the numerator of the net discrimination percentage is reduced, which reinforces their practice of boosting the denominator of that percentage by including 'neither invited' as 'equal treatment'. We cannot help wondering what Mark Twain would make of all this.

Table 1

Comparative Results for the Race Discrimination Tests in the UK

Study	Year of test/ Location	Minority	Neither invited	Usable tests (1)	Equal treatment (2)	Discrimination against minority (3) No.	Discrimination against white (4) No.	Net Discrimination (4) − (3) [(4) − (3)]/(1) No.	%
Brown and Gay (1985) – RA†	1984/5								
In-person, telephone	Birmingham	Asian/W.I‡	32§	68	48	18	2	16	24.0***
Written	London Manchester	Asian/W.I	199	267	144	102	21	81	30.0***
Daniel (1968) – RA	1966								
In-person	All major regions of England	Asian/W.I	10	30	3	27	0	27	90.0***
In-person		Hungarian		30	17	13	0	13	43.3***
In-person, employment agencies		Asian/W.I		15	4	11	0	11	73.3***
Esmail and Everington (1993) – U†	1992								
Written	England	Asian	11	12	6	6	0	6	50.0**
Esmail and Everington (1997) – U	1997								
Written	England	Asian	21	29	15	11	3	8	27.6*
Firth (1981) – U	1977/8								
Written	England	Asian	41	241	122	118	1	117	48.0***
		W.I	38	244	132	108	4	104	42.3***
		Australian	37	245	206	34	5	29	12.0***
		French	37	245	187	53	5	48	19.6***
		African	35	247	143	97	7	90	36.4***
Hubbuck and Carter (1980) – AG†	1977/9								
Written	Nottingham	Asian	58	103	48	49	6	43	42.0***
		W.I	58	103	49	49	5	44	43.0***
Jowell and Prescott-Clarke (1970) – RA	1969								
Written	Regions of:	Asian	6	26	11	14	1	13	50.0***
	Birmingham	W.I.	5	27	22	4	1	3	11.0

Table 1
Continued

Study	Year of test/ Location	Minority	Neither invited	Usable tests (1)	Equal treatment (2)	Discrimination against minority (3) No.	Discrimination against white (4) No.	Net Discrimination (4) − (3) [(4) − (3)]/(1) No.	%
	Leicester	Australian	5	27	26	1	0	1	4.0
	London	Cypriot	6	26	21	4	1	3	11.0
McIntosh and Smith (1974) – RA	1973								
In-person, telephone	Birmingham	Asian /W.I.	56	146	93	53	3	50	34.3***
In-person, telephone	London	Greek	25	32	29	3	0	3	9.0
Written		Asian /W.I.	n.a.‖	234	136	84	14	70	30.0***
Written		Italian	n.a.‖	71	52	13	6	7	10.0

Note: * significant at the 0.05 level; ** significant at the 0.01 level; *** significant at the 0.001 level.
† In this, and subsequent tables AG, RA and U signify, respectively, advocacy group, research agency, university researcher.
‡ W.I.: West Indian.
§ Exact figure not given, Brown and Gay (1985, pp. 12–3).
‖ McIntosh and Smith report only that 31 tests were not completed.

Third, we must consider the interpretation of those occasions when the 'majority' candidate is rejected, but the 'minority' candidate is offered an interview or job. There will always be some random element in hiring decisions, particularly where large applicant pools are being processed, and we agree with Heckman and Siegelman that 'there is no reason to imagine that every instance of differential treatment constitutes discrimination', and that 'discrimination exists whenever two testers in a matched pair are treated differently in the aggregate or on average' (1993, p. 198). Therefore we recommend that a measure of *net discrimination* be arrived at by deducting occasions of 'minority-only offered job' from occasions of 'majority-only offered a job'. This is the approach first advocated by McIntosh and Smith (1974, p. 13). It has subsequently been adopted by the ILO (Bovenkerk, 1992, pp. 26, 31) and Riach and Rich (1987, 1991), and recommended by Heckman and Siegelman (1993).

If the two categories are added it does provide evidence of the incidence of differential treatment and randomness in the labour market. For example consider the results for Australian payroll clerks and Austrian accountants in Table 5. In neither case is any sexual discrimination apparent, but differential treatment occurs in nearly 50% of cases for Austrian accountants and 33% of cases for Australian payroll clerks. This extent of randomness in the hiring process is inconsequential in such occupations, where there is a regular and substantial flow of jobs through the labour market, but it will not be inconsequential for conductors (or would-be conductors) of symphony orchestras, nor for captains (or would-be captains) of cruise liners. This statistic of 'total differential treatment' is an incidental by-product of the tests for employment discrimination, and is available for the reader's assessment, provided there is the full disclosure of data which we strongly recommend.

1.3. *The Tests for Racial Discrimination*

The published studies have not reported their results in a consistent format and many do not report in the detail we recommended above in Section 1.2, therefore it has been necessary to adjust some of the test results so that a comparison can be made. We report the results for employment discrimination wherever possible consistently with the above discussion, and according to the method used by McIntosh and Smith in 1973/4 (and adopted by many studies since; most notably the ILO studies of the mid 1990s). If both applicants were invited to job interview (or sent an application form in some of the studies) this represents a case of no discrimination, or 'equal treatment'. If only one applicant was invited to interview this represents a case of discrimination.

Table 1 shows the British results. The technique of making carefully matched pairs of job applications to test for discrimination in employment originated in Britain. In 1966, Political and Economic Planning designed a major study to assess the extent of racial discrimination in areas not covered by the existing Race Relations Act of 1965 – employment, housing, credit and financial services (Daniel, 1968). Surveys of immigrants, employers and employment agencies were complemented with personal testing, using professional actors, in six major regions of

Britain. A three-way match of a single tester team was used by Daniel (black minority, white minority, white national), to determine whether any discrimination found was due to colour or to national origin. The testers were sent to employment agencies and employers to test for discrimination in the labour market. The applicants were always sent in the following order: black minority first, white minority second, white national third. In particular, they were sent to firms identified in the survey of immigrants as having discriminated. Also they were sent to employers who claimed they employed blacks and did not discriminate.

Jowell and Prescott-Clarke (1970) tested for discrimination in white-collar jobs in four major regions of England. The importance of their study was that it introduced the use of written applications. Two letters of application were sent to each vacancy tested, one always from a British-born white, the other one from either an Asian (Indian/Pakistani), Australian, Cypriot, or West Indian. This enabled the researchers to distinguish 'colour' from 'foreignness'. McIntosh and Smith (1974) conducted a follow-up study for PEP in 1973, once again using professional actors for in-person approaches and telephone tests as well as sending written applications. Like the earlier PEP study, they tested to assess whether any discrimination found was due to colour or national origin. They sent matched pairs of letters, or testers, with one applicant always being a white British person, and the other one being either a West Indian, Indian, Pakistani or Italian (or Greek, in the case of the testers). For the personal approaches they created 28 pairs of testers: 8 in Birmingham, 20 in London. To ensure that the employer could correctly identify the race/ethnicity of the testers over the telephone, the testers gave their name first and then spoke with an accent. The minority tester always made the first approach. For both skilled and unskilled occupations, they applied to advertised jobs in newspapers or sent testers to factories to inquire whether jobs were available. Hubbuck and Carter (1980) and Brown and Gay (1985) used a three-way test with Asian, West Indian and white national names in written applications. In case the name used for the West Indian applicant failed to identify their ethnic origin, any doubt was removed by explaining that their early education had been completed in Jamaica (Hubbuck and Carter, 1980, p. 31; Brown and Gay, 1985, p. 9). Hubbuck and Carter and Brown and Gay also conducted a small number of telephone tests, each using 3 testers (the minority tester always phoning first), because this was the only way to test for skilled manual jobs where advertisements stipulated telephone responses. Firth (1981) applied to five types of accounting jobs in England: these being articled clerks (which required university qualifications), unqualified personnel (entry-level jobs which required no formal higher education), qualified accountants for professional firms, qualified accountants for industry and qualified accountants for financial institutions. He sent seven applications on each test, one each from an English, African, Australian, French, Indian, Pakistani and West Indian (Firth, 1981, p. 266–70). Esmail and Everington (1993, 1997) sent curriculum vitae in response to advertised medical positions in British hospitals, testing for discrimination against Asian doctors who were British trained. The above tests have covered Birmingham, Derby, Greater London, Leeds, Leicester, Manchester, Nottingham, Reading, Slough, Windsor and Wolverhampton. The occupations tested have been

accountants, clerks, hospital consultants, sales representatives, secretaries, shop assistants and, in the skilled manual area, bricklayer, carpenter, electrician, motor mechanic, painter, panel beater, plasterer, plumber, sheet metal worker and toolmaker. All the letter tests have used male applicant pairs when applying to male-dominated jobs and female applicant pairs when applying to female-dominated jobs.

The overwhelming majority of the tests have paired West Indian or Asian minorities, termed black applicants, with a British white applicant; in some cases in Table 1 it was not possible to report the West Indian and Asian paired results separately. In seven of these eight studies the level of net discrimination against black applicants was found to be equal to or greater than 30% (the Asian and West Indian results combined are 30% for the Jowell and Prescott-Clarke study). That is, over the period from the mid-1960s to the late 1980s, in nearly one in three job 'observations' the black applicant was denied an interview on the basis of race, and race alone. Firth (1981), Jowell and Prescott-Clarke (1970) and McIntosh and Smith (1974) found that the coloured immigrant group experienced a greater incidence of discrimination than the white immigrant groups.

The researchers whose studies are reported in Table 1 (and subsequent tables) have used a variety of methods to test for the statistical significance of their findings. So, for comparative purposes, and to get uniformity of treatment we have adopted the practice of McIntosh and Smith and have tested for statistical significance using the chi-squared test. Statistically significant levels of discrimination against the black applicant were found in all the studies conducted in England. McIntosh and Smith used an analysis of variance to check for tester variability and found it was insignificant, that is, it had no effect on the results (1974, pp. 16, 38–9). Letter type was found to be insignificant in the studies of Brown and Gay (chi-squared test, 1985, p. 17), Hubbuck and Carter (chi-squared test, 1980, p. 50), Jowell and Prescott-Clarke (analysis of variance, 1970, p. 409), and McIntosh and Smith (chi-squared test, 1974, pp. 53–4).

The ILO experiments were designed so that all stages of the hiring process could be tested and to ensure comparability of the results across the participating countries (Bovenkerk, 1992). All the testers were male, except for a small number of extra tests, conducted in the Netherlands, which used females. Most testers were university undergraduates. Four pairs of testers were created from two minority and two majority applicants in each country except Spain where four teams were configured in each region due to regional language differences. Testers first applied over the telephone for advertised jobs, and, if invited, attended an interview. The minority applicant always made the first phone call. The test was stopped as soon as one tester was rejected (Bovenkerk, 1992, p. 25). A number of written application tests were also conducted. Testing was conducted in major regions of those European countries participating in the project: Belgium (Brussels, Flanders and Walloonia); Germany (Berlin, Rhine-Ruhr); the Netherlands (Amsterdam, Rotterdam, Utrecht) and Spain (Barcelona, Madrid, Malaga). A variety of jobs in sales, hotels, restaurants, offices, professional and blue-collar areas were tested. The responses in these studies have been recorded and categorised according to our full disclosure recommendations. The aggregate results for each country are

reported in Table 2. In all countries significant discrimination against racial minorities was found at the initial hiring stage. Discrimination at the initial stage of hiring accounted for nearly 90% of the total level of net discrimination recorded in each country, except in Germany where it accounted for approximately two-thirds. Discrimination across both stages of the hiring process indicated that, in at least one in three observations, the minority job-seeker would have been rejected for a job in any of these countries. In all these studies, the levels of net discrimination found were statistically significant; the majority at the 0.001 level. All the researchers used chi-squared analysis to check whether there was any impact on the observations from the different pairs of testers. Bovenkerk (1992, p. 31), recommended that these be run after half the tests had been completed, so that, if any bias was found it could be corrected. No effect from individuals or pairs of testers was found in any of the studies. All the studies published the observations by tester pairs (and letter type) as well as the validity checks (Bovenkerk *et al.*, 1995, pp. 12–3; de Prada *et al.*, 1996, p. 47; Goldberg *et al.*, 1996, p. 20–21; Smeesters and Nayer, 1998, pp. 42–3).

In the United States, major field experiments of employment have been conducted by the Urban Institute in Chicago and San Diego, testing for differential treatment of Hispanics (Cross *et al.*, 1990), and in Washington DC and Chicago, testing for differential treatment of African–Americans (Turner *et al.*, 1991). The Fair Employment Council of Washington (FEC) has also conducted a large number of audits, testing for differential treatment of African–Americans and Hispanics in Washington DC (Bendick *et al.*, 1991, 1994). These studies used multiple pairs of matched testers: the UI used 10 and 8 teams for, respectively, their African–American and Hispanic tests; the FEC used 6 teams for their African–American tests and 2 teams for their Hispanic tests. The UI used only male pairs while the FEC used an equal number of female and male pairs. The testers recruited were mainly university and college undergraduates in their early twenties. In contrast to the UK and ILO studies, the UI and FEC rotated the first approach to the employer between the testers in a pair so that 50% of the time the minority tester made the first approach and 50% of the time the majority tester made the first approach. With the exception of the FEC Anglo/Hispanic tests, these studies tested whether there were differences in outcomes at any of the three stages in the hiring process. First, in success in obtaining the opportunity to apply; that is, in being provided with an application form: second, in success in obtaining an invitation to interview: third in success in being offered employment (Turner *et al.*, 1991, p. 31).

Often the first contact with an employer in these tests was over the telephone, with the tester endeavouring to obtain a form on which to submit a job application. The UI Anglo/Hispanic study arranged for the Hispanic testers to have accents, which could be detected fairly easily over the phone. However it was not possible to signal race in this manner over the phone in the African–American/White tests. So, the researchers felt that the first opportunity that the employer had to discriminate in these tests was at the second stage, after an application form had been collected and when the employer was considering which applicants to invite to interview. A variety of jobs in sales, hotels, restaurants, office and blue-collar areas were tested.

Table 2

Comparative Results for the Race Discrimination Tests of the ILO Studies

Country/Study	Year of test/Location	Minority	Neither invited	Usable tests (1)	Equal treatment (2)	Discrimination against minority (3) No.	Discrimination against white (4) No.	Net Discrimination (4) − (3) [(4) − (3)]/(1) No.	%
Belgium									
Smeeters and Nayer (1998) – U	1995/7								
Telephone – interview offer	Brussels	Moroccan/m†	243	394	99	247	48	199	50.5***
In-person – job offer	Flanders		37	62	37	19	6	13	21.0**
	Wallonia								
Germany									
Goldberg *et al.* (1996) – U	1993/4								
Telephone – interview offer	Berlin	Turkish/m	158	175	142	33	0	33	18.9***
	Rhine-Ruhr								
Netherlands									
Bovenkerk *et al.* (1995) – U	1993/4								
Telephone – interview offer	Amsterdam	Moroccan/m	151	126	62	60	4	56	44.4***
In-person – job offer	Rotterdam		12	8	0	8	0	8	100.0***
	Utrecht								
Telephone – interview offer		Surinamese/m	47	81	32	45	4	41	50.6***
Written			133	157	79	53	25	28	17.8***
Telephone – interview offer		Surinamese/f	32	83	41	39	3	36	43.4***
Written			38	78	52	18	8	10	12.8*
Telephone – interview offer		Moroccan/f	29	77	33	38	6	32	41.6***
Spain									
de Prada *et al.* (1996) – RA	1994/5								
Telephone – interview offer	Barcelona	Moroccan/m	261	268	112	141	15	126	47.0***
In-person – job offer	Madrid		25	26	9	14	3	11	42.3**
	Malaga								

Note: * significant at the 0.05 level; ** significant at the 0.01 level; *** significant at the 0.001 level.
† m – males; f – females.

Unlike the ILO method of stopping the test as soon as one of the pair had been rejected, the UI and FEC multi-stage tests involved individual testers proceeding as far as possible in the hiring process, even if their 'pair' had been rejected at an earlier stage. In particular, at the job offer stage, a number of the audits involved only one tester going to interview. In order to ensure reportage in Table 3 which is consistent with the British and ILO experiments discussed above, and reported in Table 1 and 2, we have used the report prepared by Bendick for the ILO as the source of our data for the FEC tests and as the main source for the UI White/Black tests (Bendick, 1996, p. 18). Bendick's task for the ILO was to provide data classified in the form recommended above, which enables 'neither invited to interview'/'neither offered the job' to be classified as a non-observation and deleted from the denominator of the net discrimination percentage. The exceptions were the data for the UI Anglo/Hispanic tests conducted by Cross *et al.* (1990) taken from Kenney and Wissoker (1994, p. 676), and the data for the job offer stage of the UI White/Black tests which were taken from Turner *et al.* (1991, p. 40). The reason for these exceptions is that we wish to confine the reporting of results, consistent with the ILO experiments, to interviews attended by *matched* pairs.

In the UI Anglo/Hispanic study, the level of discrimination encountered by Hispanics at the 'interview offer' stage, was twice that encountered when jobs were offered. However, in the UI White/Black study, the level of discrimination encountered by African–Americans at the 'interview offer' stage, was three-quarters that encountered when jobs were offered. Net discrimination recorded against Hispanics, in both the UI and FEC tests, was always at least 25%. Lower levels of net discrimination were recorded against African–Americans, again, in both the UI and FEC tests (at least 10%). All the above studies found statistically significant levels of discrimination against Hispanics. In the case of the African–Americans tests, statistically significant levels of discrimination were found at the 'interview offer' stage in the UI tests and at the 'job offer' stage in the FEC tests. Heckman and Seigelman tested for the impact of the pairs of testers on the observations of differential treatment for each of the UI studies. Tests of homogeneity did find homogeneity across all pairs of testers with the exception of the Black/White pairs in Chicago (1993, p. 201). Tests of symmetry indicated that little, if any, of the difference in treatment in hiring was due to the testers themselves, with the exception, again, of the audit pairs in the Black/White Chicago tests (p. 205).

James and DelCastillo (1992) used matched pairs of testers in Anglo/Hispanic and Black/White audits for job offers in Denver. James and DelCastillo used the same methodology as the UI for the conduct of their tests, sending a tester to interview if they were offered one, even if their 'pair' had been rejected. Even though their published data enable 'neither invited to interview' to be distinguished from 'neither offered the job' it is not possible to confine the reporting of results to interviews attended by matched pairs, nor to report the results of the two-stage audits separately for 'invitation to interview' and 'job offer'. For these reasons we have not been able to include their findings in Table 3. Moreover James and DelCastillo's use of bonus payments for the testers may have introduced some important differences in motivation. In their experiment both testers in a pair received a bonus when at least one of them was successful at any stage of the hiring

Table 3

Comparative Results for the Race Discrimination Tests in the United States

Study	Year of test/ Location	Minority	Neither invited	Usable tests (1)	Equal treatment (2)	Discrimination against minority (3) No.	Discrimination against white (4) No.	Net Discrimination (4) − (3) [(4) − (3)]/(1) No.	%
Bendick *et al.* (1994)† – AG	1990/1								
In-person – interview offer	Washington	Black	8	141	125	10	6	4	2.8
In-person – job offer			101	24	5	18	1	17	70.8***
Bendick *et al.* (1991)† – AG	1992‡								
In-person – interview offer	Washington	Hispanic	99	183	137	46	0	46	25.1***
Cross *et al.* (1990)§ – RA	1989								
In-person – interview offer	Chicago	Hispanic	103	257	158	77	22	55	21.4***
In-person – job offer	San Diego		39	101	62	25	14	11	10.9*
Turner *et al.* (1991)‖ – RA	1990								
In-person – interview offer	Chicago	Black	211	265	215	35	15	20	7.6**
In-person – job offer	Washington		129	89	62	18	9	9	10.1***
Nunes and Seligman (1999) – RA	1998								
In-person – job offer	San Francisco	Black	0	45	12	25	8	17	37.8***

Note: * significant at the 0.05 level; ** significant at the 0.01 level; *** significant at the 0.001 level.

† For comparative purposes the results are reported as compiled in the ILO paper on the US by Bendick (1996), Table 2, p. 18.

‡ Bendick *et al.* (1991), report that the tests were conducted in 1992.

§ As reported in Kenney and Wissoker (1994), Table 2, page 676.

‖ The data for the interview stage are reported in the ILO paper on the US by Bendick (1996), Table 2, p. 18. The data for the job offer stage are for the audits with both partners remaining and are reported in Turner *et al.* (1991), Table 4.3, p. 40. The number of 62 for equal treatment is taken from Bendick (1996), Table 2, p. 18, which is consistent with Heckman (1998), Table 1, p. 105.

process. The testers could receive up to three bonuses, one for submitting a job application, one for obtaining an interview, and one for receiving a job offer. This method of payment could have affected the relative incentive of testers to succeed, particularly if their 'pair' had already achieved an interview or a job offer. So the results of the Denver study should be treated with some caution (Bendick *et al.*, 1994, pp. 45–6, fn. 11; Mincy, 1993, p. 176; Zimmermann, 1993, p. 410).

Nunes and Seligman (1999) sent three matched pairs of testers, consisting of one Caucasian and one African–American, to employment agencies in San Francisco. They found less favourable treatment of African–American applicants in 38% of the tests. In this case less favourable treatment involved quality and pay of jobs rather than denial of interview. All testers, except one, were granted interviews.

Tests for racial discrimination have also been conducted in Australia over the period 1986–8 (Riach and Rich, 1991), Canada in 1984 (Henry and Ginzberg, 1985) and France over the period 1976–7 (Bovenkerk *et al.*, 1979). The Canadian study conducted in Toronto, involved approaches to employers, over the telephone and in-person, using four teams, two female and two male. In the case of the in-person tests one tester was always a White Canadian and the other a West Indian. Two curriculum vitae for each job type were prepared and were rotated evenly among the testers in a team. In the telephone applications which used three testers, one tester was always a White Canadian one a West Indian and a third was Italian. In these telephone applications, the tester's race/ethnicity was conveyed by their name and by the use of accents. This approach was successfully checked in a preliminary test (Henry and Ginzberg, 1985, pp. 43–4). This design was used so as to imitate the British tests which had investigated for differential treatment based on colour as compared to 'foreignness'. As Henry and Ginzberg wished to test senior job positions, as well as semi and unskilled jobs, one of the two teams for each sex consisted of testers in their thirties. Professional actors were hired for these older tester positions, because it was felt that they would have the skills to sustain the more difficult roles required for the senior job positions. University and high school students were hired for the younger tester positions (Henry and Ginzberg, 1985, p. 19). Over the course of the three and a half months of the study they used fourteen testers to create different combinations of teams. The Black applicant made the first approach in all tests. They applied to sales and professional positions such as, accountants, office manager, executive secretary, salesperson, and secretary. The Australian study conducted in Melbourne used written approaches and covered three occupations; clerk, sales representative and secretary. Two applications were sent to each selected vacancy; one always from an Australian, the other one from either a Vietnamese or a Greek. Written approaches were also used in the French study. Personal approaches could not be used because it was not possible to signal race (Antillian) over the phone. Two applications one from a Frenchman, the other from an Antillian, with photos attached (a normal practice), were sent to non-manual jobs which were advertised in newspapers.

Racial discrimination was found in all these countries (see Table 4). In the Canadian in-person tests a level of net discrimination against the Black applicant

of 39% was found. In the case of the telephone tests, we are only able to report that of the 237 tests conducted, the white Canadian received 206 positive responses, the white immigrant received 154 positive responses and the black Canadian received the least positive responses – 123. In Australia a much higher level of net discrimination was recorded against Vietnamese (27.4%) than against Greeks (8.8%). A high level of net discrimination was recorded against Antillians in France (67%). In all these studies the levels of net discrimination found were statistically significant. No effect from letter type was found in the Australian study.

The results of the racial discrimination tests have extended over a period of thirty years and nine countries, in Europe, North America and the Pacific; all are members of the OECD. The minority groups include black, Asian, Arab, Turkish and other white non-nationals. The extent of discrimination varies temporally, spatially and between the various minority groups. What is most significant though, is that, with the exception of the studies in the United States, the rate of net discrimination recorded against blacks, Asians, and Arabs has never been less than 25%. In view of the number of studies involved and their geographical extent this is compelling evidence of enduring and pervasive racial discrimination in employment. Before we turn to consider the implications of these findings for economic theory it is necessary to speculate about the reasons for the lower level of racial discrimination recorded in the US labour market. In the United States, the 1964 Civil Rights Act established the Equal Employment Opportunity Commission, and in 1965, Executive Order 11246 established the Office of Federal Contract Compliance which implemented affirmative action programmes. One obvious explanation then, is that this intensive government activity dating back to 1964 has achieved some success in diminishing employment discrimination against African–Americans. Another factor is the location of the tests which influences results. US cities have different industrial sectors, different local labour market conditions, and different local minority concentrations. Bendick (1996, p. 27) notes that rigorous evaluation of this phenomenon is not possible 'until additional tests are conducted in a broader range of locations'.

These employment experiments have not been designed to distinguish between the various hypotheses which have been promulgated to account for discrimination, but the pattern of results does enable some tentative speculation, in the manner of Neumark *et al.* (1996, pp. 936–7). The 'statistical' theory (Phelps, 1972) postulates differences, on average, between racial (sexual) categories, in their employment characteristics. Consequently race or sex is used as a cost-minimising screening device – 'Skin color or sex is taken as a proxy for relevant data not sampled' (Phelps, 1972, p. 659). Given the very diverse cultural, social, religious and educational backgrounds of the various Asian, Arab and African-descendant groups who have encountered the employment discrimination in these studies, it would be a superficial generalisation to hold that the recorded disinclination to hire them arises because the various 'indigenous' white populations have, on average, superior employment characteristics. The common characteristic of Asians, Arabs and African-descendant groups is that

Table 4

Comparative Results for the Other Race Discrimination Tests

Study	Year of test/ Location	Minority	Neither invited	Usable tests (1)	Equal treatment (2)	Discrimination against minority (3) No.	Discrimination against white (4) No.	Net Discrimination (4) − (3) [(4) − (3)]/(1) No.	%
Australia									
Riach and Rich (1991) – U	1986–88								
Written	Melbourne	Vietnamese	362	157	96	49	9	43	27.4***
		Greek	292	170	135	25	10	15	8.8**
Canada									
Henry and Ginzberg (1985) – RA	1984								
In-person – job offer	Toronto	West Indian	155	46	10	27	9	18	39.1***
France									
Bovenkerk *et al.* (1979) – U	1976–77								
Written		Antillian	415	267	55	195	17	178	66.7***

Note: ** significant at the 0.01 level; *** significant at the 0.001 level.

Table 5

Comparative Results for the Sex Discrimination Tests

Study	Year/Location of test	Occupation	Neither invited	Usable tests (1)	Equal treatment (2)	Discrimination against female (3) No.	Discrimination against male (4) No.	Net Discrimination (4) − (3) [(4) − (3)]/(1) No.	%
Australia									
Riach and Rich (1987) – U	1983-86								
Written	Melbourne	Analyst prog.	59	93	70	17	6	11	11.8*
		Computer operator	50	49	35	8	6	2	4.1
		Computer prog.	44	71	53	7	11	−4	−5.6†
		Gardener	86	62	42	15	5	10	16.1*
		I.R. officer	56	38	26	5	7	−2	−5.2
		Man. Accountant	103	108	80	18	10	8	7.4
		Payroll clerk	86	86	57	14	15	−1	−1.2
Austria									
Weichselbaumer (2000) – U	1998/9	Accountant							
Written	Vienna	male/masc. female‡	69	80	38	21	21	0	0
		male/fem. female‡	37	112	63	22	27	−5	−4.5†
		m.female/f. female‡	73	76	47	12	17	−5	−6.6
		Computer prog.							
		male/masc. female	8	80	67	5	8	−3	−3.8*
		male/fem. female	7	81	62	10	9	1	1.2**
		m. female/f. female	9	79	67	8	4	4	5.1
		Network technician							
		male/masc. female	24	93	66	19	8	11	11.8
		male/fem. female	26	91	62	23	6	17	18.7
		m. female/f. female	35	82	60	14	8	6	7.3
		Secretary							
		male/masc. female	59	64	18	7	39	−32	−50.0**
		male/fem. female	61	62	17	8	37	−29	−46.8**
		m. female/f. female	55	60	43	14	11	3	0.1
US									
Levinson 1975 – U	1974								
Telephone	Atlanta	Male-dominated	51§	95	54	41	0	41	43.2***
		Female-dominated	35§	74	26	0	48	48	64.9***

Table 5
Continued

Study	Year/Location of test	Occupation	Neither invited	Usable tests (1)	Equal treatment (2)	Discrimination against female (3) No.	Discrimination against male (4) No.	Net Discrimination (4) − (3) No.	Net Discrimination [(4) − (3)]/(1) %
Neumark *et al.* (1996) – U	1994	Waitress							
In-person	Philadelphia								
Restaurant – interview offer		High price	6	17	3	11	3	8	47.0**
		Medium price	6	15	7	6	2	4	26.7
		Low price	11	10	2	2	6	−4	−40.0†
Restaurant – job offer		High price	11	12	1	10	1	9	75.0***
		Medium price	9	12	4	6	2	4	33.3
		Low price	13	8	2	0	6	−6	−75.0**†
Nunes and Seligman (2000) – RA	2000								
In-person – initiating enquiry	San Francisco	Auto service	0	40	13	19	8	11	27.5**
In-person – when jobs available			0	20	5	12	3	9	45.0**

Note: * significant at the 0.05 level; ** significant at the 0.01 level; *** significant at the 0.001 level.
† A negative value indicates discrimination against the male applicant.
‡ masc. – masculine; fem.- feminine; m. female – masculine female; f. female – feminine female.
§ Levinson classified some audits as 'not discernable'. 'Ambiguous' results have not been included in usable tests: females applying to male-dominated = 46; males applying to female-dominated = 24.

they are not white. The seven countries recording employment discrimination of upwards of 25% against these three broad groups all have a predominantly white population, and experienced substantial non-white immigration during the last 50 years. White immigrant groups (for example, Italians, Greeks) encountered discrimination in some studies, but never at a level comparable to non-whites. These results are consistent with Darity and Nembhard (2000), who found persistent discrimination against men of colour in their study of relative income and earnings in twelve countries. The findings of these field experiments are more consistent with the majority white populations having a *general* 'distaste' (Becker, 1971), or 'social custom' (Akerlof, 1980), which motivates employers to discriminate against non-white applicants. But we stress that field experiments have not, to date, been designed so as to enable any firm conclusion about the nature of discrimination.

1.4. *The Tests for Sex Discrimination*

Table 5 provides details of five experiments. In the United States, Levinson (1975) conducted telephone tests in Atlanta of occupations in which either male or female employees dominated, such as child-care workers, hairdressers, motor mechanics, and secretaries. Approximately 50 testers were used in pairs, and the 'sex-inappropriate' tester always rang first. A minority of the tests could not be classified either because the employer gave an ambiguous response or, simply, did not provide enough information. 'Clear-cut' discrimination, defined as a clear restriction or elimination of an applicant due to their sex, was found against women in male-dominated occupations (43.2%) and against men in female-dominated occupations (64.9%). Over the period March to May 1994 Neumark *et al.* (1996) sent paired testers to drop off curriculum vitae to restaurants in Philadelphia, leaving these with employees and asking that they be forwarded to the employer/manager. In effect they were written tests, but hand-delivered, rather than mailed. Like the earlier British study of McIntosh and Smith and the ILO studies, the testers (two males and two females) were rotated to create (four) teams. The curriculum vitae and the 'first approach' were rotated among the job applicants. They found net discrimination, (47%) against females in high price (higher pay) restaurants but against men (40%) in low price (lower pay) restaurants. In Neumark *et al.*'s study, the net discrimination against females in high price restaurants was statistically significant for both interview and job offers. The net discrimination against males in low price restaurants was statistically significant only for the job offer stage of hiring. They found no effects on the results of the audits for curriculum vitae, individual tester or pair of testers (Neumark *et al.*, 1996, pp. 928–30).

In 2000, Nunes and Seligman (2000) conducted 40 in-person tests of auto services jobs in San Francisco. They sent two pairs of testers to 20 auto shops with the female tester always making the first approach. They found net discrimination against females of 27.5%. A much higher level of net discrimination against females (45%) was recorded in auto shops that actually had vacancies advertised at the time of the approach.

Firth (1982), sent written applications in 1978 to British accountant advertisements and found a statistically significant lower success rate of females in the two higher status, career areas of 'qualified accountant' and 'senior jobs in commerce and financial institutions' (p. 897). He also tested for the impact of colour and children on the success rate of females. Coloured females with children had the lowest success rate of all the job applicants, in getting an invitation to interview (p. 896). The published data for this study are not reported by matched pairs or in the classifications consistent with the British and ILO race experiments discussed above, therefore it is not possible to report his findings in detail. Riach and Rich (1987) sent written applications to seven occupations in Melbourne during 1983 to 1986. They found statistically significant net discrimination against women in the occupations computer analyst programmer (11.8%) and gardener (16.1%). During 1998 and 1999, Weichselbaumer (2000), sent written applications in Vienna to two sex-integrated occupations (computer programmer, accountant), one female-dominated occupation (secretary) and one male-dominated occupation (network technician). Three job applicants were invented, a male, a 'masculine' female and a 'feminine' female. Photographs of the applicants were attached to the job applications (a normal procedure in Austria) and independent tests of the curriculum vitae indicated that the photographs were effective in conveying the different female personality types (Weichselbaumer, 2000, pp. 19–20). The hypothesis being tested was that females with 'masculine' traits would be treated like males, for male-dominated jobs while females with 'feminine' traits would be considered inappropriate for these jobs. She found net discrimination against both types of women in the male-dominated occupation (at least 11.8%) and against men in the female-dominated occupation (at least 46.8%). No statistically significant difference in treatment was found between the two female types in any of the occupations.

In male-dominated occupations, such as motor mechanic and network technician, and in the female-dominated occupations such as secretary, the results have always been statistically significant. Moreover some of the studies have found statistically significant discrimination in 'sex-integrated occupations', such as accountancy. When we draw the six studies together two significant regularities emerge in the data. First, women are particularly prone to encounter discrimination in higher status and/or hierarchically senior jobs. Riach and Rich (1987) combined the data for the two most hierarchically senior occupations in their study (computer analyst programmer and management accountant) and found statistically significant discrimination against women. Neumark *et al.* found discrimination against females in high price, higher pay, restaurants, but that they were favoured in low price, lower pay, restaurants. Firth (1982) found a lower success rate of females in the two highest status accountants' job areas – 'qualified accountant' and 'senior jobs in commerce and financial institutions'. Second, whilst 'integrated' occupations, such as computer programmer and payroll clerk, sometimes recorded an absence of discrimination, when sex-stereotyped occupations were investigated, significant discrimination was always recorded. In the traditional female occupation secretary, Levinson and Weichselbaumer found substantial discrimination against men. Our current study of sex discrimination in England is producing identical results for the occupation of secretary. Likewise

Levinson, and Nunes and Seligman recorded substantial discrimination against women in the traditional male occupation of motor vehicle mechanic. In particular, discrimination against men in 'female' occupations was always much higher than that against women in 'male' occupations.

In an interesting and related study, Goldin and Rouse (2000) used naturally occurring data, rather than data generated by experiments, to test for sex discrimination in the hiring of musicians by orchestras. They used the hiring outcomes when orchestras did, and did not, use a screen to hide the musician auditioning, together with other relevant information, to estimate a model explaining the probability of an individual advancing in an audition (p. 726). They found persuasive evidence that a screen increased the probability that a female would be hired by an orchestra.

In the case of women, we would expect Phelps-style statistical discrimination to be particularly directed at their maternal role; their job tenure under suspicion because of possible pregnancy and their reliability suspect because of child-care responsibilities. Discrimination of such a statistical nature would most likely be associated with occupations which have substantial fixed costs of employment, thereby deterring the employment of groups with higher turnover rates. Neumark *et al.* (1996) so speculate about their finding of discrimination against women in high-price restaurants, but conclude that such 'statistical' considerations were unimportant in explaining this result (p. 937). The occupational pattern of discrimination in these studies is not consistent with this statistical discrimination hypothesis. The Riach and Rich (1987) and the Weichselbaumer studies both found that discrimination against women was absent in computer programmer jobs, and Riach and Rich (1987) also found it was absent in industrial relations officer jobs. On the other hand Riach and Rich (1987) found discrimination against females in computer analyst programmer jobs and Weichselbaumer (2000) found discrimination against females in network technician jobs. Such results are not consistent with a sexually differential incidence of fixed costs generating statistical discrimination. Female computer programmers and industrial relations officers are not dissimilar to the latter two occupations in the need for on-the-job training and experience and presumably they are just as susceptible to pregnancy.

The key regularities found in these tests of sex discrimination were a particular incidence against women in the more senior, high status/pay jobs, and against both sexes when they applied for an occupation which was 'stereotyped' for the other sex. This is consistent with the hypothesis that many in society still identify appropriate roles for men and women: in particular, deeming women to be inappropriate in jobs associated with high status or physical effort: the supportive and decorative role of secretary being deemed inappropriate for men, whilst the dirty and physical-demanding nature of motor mechanics being deemed unsuitable for women.

1.5. *The Test for Age Discrimination*

One test for discrimination on the basis of age is reported in Table 6. Bendick *et al.* (1999) used four pairs of applicants, three male and one female, to test for age

Table 6

Comparative Results for the Age and Disability Discrimination Tests

Study	Year of test Location	Minority	Neither invited	Usable tests (1)	Equal treatment (2)	Discrimination against minority (3) No.	Discrimination against majority (4) No.	Net Discrimination (4) − (3)[(4) − (3)]/(1) No.	Net Discrimination (4) − (3)[(4) − (3)]/(1) %
Age tests									
Bendick *et al.* (1999) – AG	1995/6								
In-person, telephone	Washington	Older	38	102	70	32	0	32	31.4***
Disability tests									
England									
Fry (1986) – AG	1986								
Written	London	Disabled	n.a.	93	52	38	3	35	37.6***
Graham *et al.* (1992) – AG	1990								
Written	London	Disabled	103	94	51	37	6	31	33.0***
Netherlands									
Gras *et al.* (1996) – RA	1995								
Written	Amsterdam Rotterdam Utrecht	Disabled	116	154	63	64	27	37	24.0***

Note: * significant at the 0.05 level; ** significant at the 0.01 level; *** significant at the 0.001 level; n.a. not available.

discrimination in entry-level jobs for management and sales occupations in Washington DC, over the year from March 1995 to March 1996. The pairs were trained together and curriculum vitae were prepared for use in the tests. One tester, purported to be 57 years old, the other 32 years old. The testers first applied to advertised jobs either over the telephone, fax, mail or in person, depending on which procedure was specified in the job advertisement, with the older applicant making the first approach in the case of the telephone and in-person tests. To test for discrimination in the offer of jobs, the testers were sent to a small number of interviews when both, initially, had been treated equally. At the initial stage of the hiring process Bendick *et al.* found a level of net discrimination against the older job seeker of 31.4% and *no reverse discrimination*, ie, they found no cases of discrimination against the younger applicant at this stage of hiring. When the testers went to a small number of interviews, in 11 out of the 12 cases only the young person was offered the job; only one case was recorded where the older person was favoured. Thus a further 9.8% net discrimination against the older applicant was found. The majority of net discrimination (76%) occurred at the initial stage of the hiring process. The level of net discrimination found against older job-seekers was statistically significant. Thus, as with the ILO race studies, discrimination against older workers in the hiring process was overwhelmingly the denial to interview.

Age provides a particular challenge for these experiments, whose general purpose is to determine whether candidates with identical human capital, and productivity, incur differential employment responses because of some personal characteristic, such as, race, sex, sexual preference etc. One would expect older people to have some genuine human capital differences from the young. On the positive side one would expect them to have more experience; on the negative side they might have less physical stamina and be less receptive to new production techniques. Bendick *et al.* (1999) attempted to control for the experience component of human capital by way of a 25-year gap for the 57-year-olds, during which they were engaged in unrelated activities, such as military service and school teaching which provided no relevant experience for the targeted job. Both the candidates did have several years relevant experience for the targeted job (Bendick *et al.*, 1999, p. 8).

We have the following recommendations for future age experiments: first the age gap should be much smaller and the age of the older candidate should be lower. Anecdotal evidence in Australia and Britain suggests that age discrimination impacts long before the 57th birthday; particularly in the case of women. It would be sensible to test this hypothesis with a range of age levels and gaps. One which we suggest is women, newly-graduated, aged 22 and 37. The incidence of 'mature-age' female students in British universities and colleges makes this a very believable and relevant age difference to test. Second we suggest controlling for other human capital variables by including in the curriculum vitae of older candidates evidence of 'youthfulness', particularly physical stamina and adaptability. This could be involvement in veteran athletics, and expertise in the latest ICT developments. Third we suggest adopting Newman's (1978) procedure of sending *non-equivalent* curriculum vitae; for example, for 28- and 48-year-old, career-long accountants, secretaries and sales representatives. Physical stamina and flexibility should not be

too much in doubt by age 48, so any finding of discrimination against 48-year-olds, with 20 years more of relevant experience, would be a very significant and disturbing finding. On the other hand, a finding of equivalent treatment would provide re-assurance that experience compensated for any presumed diminution in stamina and flexibility. One would expect to find genuine differences in human capital between age groups, therefore it is not meaningful to control for comparability. It makes more sense to acknowledge the heterogeneity and control for the differences to be normally expected between the age groups being tested. Any differential response by employers to such realistic human capital circumstances is of far more relevance to policy makers, than the artificial situation contrived by Bendick *et al.* (1999). In conclusion we stress that the results of age experiments need to be interpreted with caution, and the alternatives carefully appraised. For this reason it is particularly important that full details of procedures and results are published.

1.6 *The Tests for Disability Discrimination*

The design of tests to detect disability discrimination is even more challenging than in the case of the foregoing tests for age discrimination. In many jobs there are clear and objective human capital and employment cost differences between the disabled and the able-bodied, and therefore differential treatment may well not be discrimination. Peter Cook's refusal to cast a 'one-legged' Dudley Moore as Tarzan cannot be construed as discrimination. Nevertheless it is important to ascertain whether the disabled do encounter discrimination when human capital and employment cost differences are either non-existent, or negligible: also the extent to which differential treatment is prompted by various disabilities. In the latter case it would be important to title any report carefully, as one investigating differential treatment.

There have been three studies, two in England (Fry, 1986; Graham *et al.* 1990), one in the Netherlands (Gras *et al.* 1996) that have used written approaches to test for discrimination on the basis of disability (see Table 6). While the Dutch study tested for various types of disability: being confined to a wheelchair, epileptic or deaf, the English studies only tested for one type of disability – an applicant with cerebal palsy who was confined to a wheelchair. The first study conducted was in England in 1986 with a follow-up in 1989–90, and both involved secretarial positions. These tests found statistically significant levels of net discrimination against the disabled applicant of roughly the same magnitude (37.6% in 1986 compared with 33% in 1990). As in other studies using written approaches, no effect was found from letter type in the English tests. The study in the Netherlands found a statistically significant level of net discrimination of 24% against the disabled applicant with no significance attaching to the type of handicap or the occupation (professionals in administrative, commercial and secretarial areas).

It is important to assess the extent of differential treatment against the extent of disability, before reaching any conclusion about a finding of discrimination. We reiterate our strong recommendation that full details of the design and conduct of experiments be published so that the reader can arrive at her/his own judgement

about the extent (if any) of discrimination. It is appropriate that tests be conducted for a variety of types and levels of disability as a basis for government policy aimed at educating employers about the disabled, and encouraging their employment.

1.7 *Labour Market Conclusions*

Field experiments in the labour market have found discrimination against Vietnamese and Greeks in Australia; West Indians, Pakistanis, Indians and Italians in Britain; West Indians in Canada; Antillians in France; Turks in Germany; Moroccans in the Netherlands, Spain and Belgium; Surinamese in the Netherlands; African–Americans and Hispanics in the United States; against males applying for female-dominated jobs and against females applying for male-dominated jobs in Australia, Austria and the United States. There is also evidence of differential treatment, (which may include an element of discrimination), of the disabled in Britain and the Netherlands. Finally, we have some preliminary indication of age discrimination in the United States.

These labour market experiments have been conducted most extensively for race and sex. The diversity of the non-white groups encountering discrimination, and the occupational pattern of the sex-based discrimination, suggests that it is unlikely to have been 'statistically' motivated. The recorded discrimination is more consistent with widespread social attitudes or tastes, to which employers feel obliged to conform: this is particularly apparent in sex-stereotyped jobs, such as secretary and motor mechanic.

Many of the studies have used multivariate analysis or chi-squared tests to determine the effect of other factors such as, firm size, occupation, industry, location (urban/rural/region), race/sex of employer and qualifications on the level of discrimination recorded (Bendick, Brown and Wall; Brown and Gay; both studies by Firth; Hubbuck and Carter; all the ILO studies; Jowell and Prescott-Clarke; McIntosh and Smith; Neumark *et al.* ; both studies by the UI; Weichselbaumer). Some have also recorded differences in treatment when the applicants have both succeeded at obtaining an interview or a job offer, such as minority applicants being treated less courteously at interviews, or being offered lower wages (see for example, both FEC studies; all the ILO tests; Levinson; the UI African–American tests).

Experiments involving personal approaches have been subject to criticism, since their outset, about the matching and motivation of testers and the possibility of unobserved variables. These criticisms, first made by Ward (1969) and more recently by Heckman (1998), were considered at length in Section 1.1 above, where we recommend written tests as a solution to the matching/motivation problem: a solution endorsed by Darity and Mason (1998, pp. 80–1). 'An objective demonstration of the quality of the matches would go a long way toward making audit pairs credible' (Heckman and Siegelman 1993, p. 271). This is precisely what correspondence tests do.

In Section 1.2 above we recommended that the appropriate measure of discrimination is – offers to the 'majority group-only', minus offers to the 'minority group-only', expressed as a percentage of all tests where one or both applicants

receive an offer. This is the procedure introduced by McIntosh and Smith (1974) and followed by the ILO.

We believe that the studies surveyed above demonstrate that labour market field experiments are a valuable research tool to complement the Blinder/Oaxaca regression analysis. The relative strength of 'audits' has been acknowledged by their foremost critics; 'The other major advantage of the audit technique is that it allows more control over the characteristics that are thought to be relevant to the employment decision than is possible in conventional ex-post regression analysis' (Heckman and Siegelman, 1993, p. 193).

We have offered recommendations above for the future development of 'age' and 'disability' studies. We have previously (Riach and Rich, 1998) made a suggestion to deal with a deficiency noted by Heckman and Seigelman, – 'Audit pair studies as currently conducted cannot distinguish between animus-based discrimination and statistical discrimination, although it is of scientific and policy interest to do so' (Heckman and Siegelman, 1993, p. 243). We have suggested deploying two pairs of applicants in a test of sex discrimination: one pair – each of whom is married, childless and aged 25: a second pair – each of whom is married, with one child of 12 and aged 37. If a disinclination to hire women is based in their biologically-determined role to bear children and/or in their socially-determined role to care for sick children we should record a higher level of discrimination against the woman in the first pair (Riach and Rich, 1998, p. 198). Such a finding would indicate the extent, if any, of statistical discrimination. One final observation is that it is strange that, to our knowledge, no test of religious discrimination has ever been conducted, given the anecdotal evidence, especially in Northern Ireland.

2. Field Experiments in the Housing Market

2.1. *The Technique and its History*

All the housing tests have been conducted using personal approaches. The tests in the United States have all involved in-person approaches. In Britain, while most tests have involved in-person approaches, there have been some telephone tests conducted for rental properties. Two testers are matched; one from the majority group, say white, the other from the minority group, say black, and they then approach the prospective landlord or real estate agents to look at suitable rental accommodation or houses/flats for sales. Some tests in Britain have involved sending three matched testers (for example, a Briton, a West Indian and a Greek) for housing inquiries. The US tests, again, have all used two matched applicants where one applicant is White and the other either African–American or Hispanic.

The methodology of using testers to test for discrimination in the housing market is essentially the same as that used for testing in the labour market. The testers differ in one characteristic such as sex, race, or ethnicity. The matched pairs are trained so that both testers in the matched pair make equivalent inquiries to the prospective landlord or real estate agent and their background personal characteristics, such as credit and tenant worthiness are equivalent. They are trained together and carefully supervised throughout the testing period.

In making the housing inquiries there must be some time delay between the testers' approaches, say one hour to three hours. It is of course possible that the property may be offered to another candidate in between the contacts made by the testers. Therefore the order of approach, to the housing agent, of the testers must be considered. Many of the tests ensure that in half of the tests the first approach is made by the minority applicant and in the other half, by the majority applicant. Given the intent of the experiments is to determine the extent of discrimination against the minority group then, when the minority tester making the second approach is discriminated against, it is necessary to determine if the property has been sold or rented, in the interim, to a genuine customer.

Discriminatory behaviour in housing markets is of a slightly different nature to that in labour markets and as such is recorded differently. The housing tests measure discrimination as 'opportunity denying' and 'opportunity diminishing' treatment. 'Opportunity denying' treatment covers acts of denial of any information. 'Opportunity diminishing' treatment refers to being told of fewer units, shown fewer units, being quoted less favourable rental terms, or given less information on financing potential (Yinger, 1995, pp. 31–3). There has been an extensive discussion on which is the more appropriate measure of discriminatory behaviour in housing: the net or gross level (Yinger, 1993, pp. 70–80; 1995, pp. 43-6). Discrimination against whites in housing tests may involve denial of information about sales and rental units in black residential neighbourhoods.

2.2. *Discrimination in the Housing Market*

The results reported in Table 7 are based on discrimination being measured as 'opportunity denying' treatment (as defined by Yinger), as the majority of tests on housing discrimination have been so reported in the United States.

In the first PEP study in 1966, three professional actors, one British, one West Indian and one Hungarian, were sent to landlords and real estate agents across England to test for discrimination in the housing market (sales and rental properties). The tests covered Birmingham, Greater London, Leeds, Leicester and Slough. In the second PEP study in 1973 a matched pair of testers, one always British, the other either a West Indian, Indian, Pakistani, or Greek, were sent to landlords and real estate agents in Birmingham, and London. Much lower levels of net discrimination were found for sales and for rental properties in 1973 than in 1967. McIntosh and Smith concluded that the sharp decrease in discrimination which was recorded between the two studies was due to the introduction of the Race Relations Act in 1968 (McIntosh and Smith, 1974, p. 19).

Many of the US housing audits have been conducted by private and public organisations seeking information about their own residential areas and have been prepared as written reports which are not publicly accessible. Galster (1990*a*) obtained information on 71 audits from these organisations and this is the source of a number of the US studies reported in Table 7. Wienk *et al.* (1979) used

Table 7

Audits Testing for Racial Discrimination in Housing Sales and Rental Accommodation

Country/Region*	Audit date	No of audits	Race/Sex	Net discrimination† (%)
Sales:				
England				
Birmingham, Greater London, Leeds, Leicester, Slough (Daniel, 1968)	1966	228	West Indian	60.1
Birmingham, London (McIntosh and Smith, 1974)	1973	137	Asian/W.I.	12.4
US				
National (Wienk *et al.*, 1979)	1977	3,264	Black	16.7‡
National (Yinger, 1993)	1989	1,081§	Black	30.0
National (Yinger, 1993)	1989	1,076§	Hispanic	23.0
Boston, MA	1981	118	Black	13.1
Grand Rapids, MI	1981/2	100	Black	24.0
Rental:				
England				
Birmingham, Greater London, Leeds, Leicester, Slough (Daniel, 1968)	1966	228	West Indian	60.1
London, Birmingham (McIntosh and Smith, 1974)	1973	41	Asian/W.I.	26.8
France				
Paris (Bovenkerk *et al.*, 1979)	1976	135	Black	31.9
US				
National (Wienk *et al.*, 1979)	1977	3,264	Black	16.3‡
National (Yinger, 1993)	1989	801§	Black	28.0
National (Yinger, 1993)	1989	787§	Hispanic	23.0
Boston, MA	1981	156	Black	24.0
Sunnyvale, CA	1981	23	Black	61.0
Redwood City, CA	1982	35	Black	69.0
Palo Alto, CA	1983	20	Black	70.0
Hayward, CA	1984/5	25	Black	20.0
Cleveland Heights, OH	1985	29	Black	14.0
Wooster, OH	1985/6	15	Black	20.0
Washington, DC	1986	280	Black	48.0
Washington, DC	1988	295	Black	28.0
Carmichael/Citrus Heights	1982	18	Hispanic	22.0
Redwood City, CA	1985	32	Hispanic	47.0
Hayward, CA	1985/6	25	Hispanic	4.0‖
Wooster (Galster and Constantine, 1991)	1985	11	Female	81.8

* Unless stated otherwise the studies are from Galster (1990*a*).
† All tests statistically significant at the 0.05 level, except‖.
‡ These figures are calculated as an average of advertised units available, number of units inspected and number of units available, as reported in Yinger (1995, p. 47).
§ Yinger (1991, p. 327); results for net discrimination from Yinger (1993, p. 21).
‖ Not statistically significant.

personal approaches to test for racial discrimination in the housing market in the first major national study in the United States: the Housing Market Practices Survey (HMPS). They conducted 3,264 tests in 1977 and found evidence of significant discrimination against African–Americans in sales (16.7%) and rental markets (16.3%) (Yinger, 1995, p. 20). The second major study in the United States, the 1989 Housing Discrimination Study (HDS), found that discrimination

against African–Americans in sales and rental markets had persisted (net levels of, respectively, 30% and 28% although the findings of the two studies are not strictly comparable because of a different audit methodology). They also found discrimination against Hispanics, recording net levels of 23% in both sales and rental markets.

Of the 20 audits which we report for the United States, covering sales and rentals, 15 found a level of net discrimination greater than 20%. All the findings were statistically significant except for a single audit which found very little net discrimination (Hayward, CA 1985/6). There is strong evidence therefore of significant and persistent discrimination in housing markets in the United States (Yinger, 1998*b*). In estimating the significance of the findings of differential treatment from audits Yinger (1986, pp. 883–4) demonstrates that using Ordinary Least Squares analysis may overstate the error term, leading to the conclusion that there is no discrimination when in fact there is. He recommends using a paired difference-of-means test to eliminate this overstatement (Yinger, 1986, pp. 884–5). He also used regression analysis to test whether other variables besides race explained the measured variation in discriminatory behaviour, such as the age of children, childless, income, the age and sex of the tester, the age and sex of the real estate agent, and aspects of the timing of the audit. He found that little else besides race explained the observed pattern of discrimination.

In France, three testers, one French, one white foreigner (Portuguese) and one Black (Antillian) were sent to real estate agents to test for discrimination in the housing market for rental properties. The tests were conducted in Paris in 1976. The study found net discrimination against the black applicants of 31.9% but found no discrimination against the Portuguese (Bovenkerk *et al.*, 1979).

Galster and Constantine (1991) conducted eleven audits, over the phone and in-person, testing for discrimination against female-headed families using three matched testers, females with and without children and a male without a child. A level of net discrimination of 82% against both types of female-headed households was found, which was statistically significant.

These audits in Britain and the United States found considerable evidence of 'steering' in housing markets – individuals being shown the same number of units but in different residential areas dependent on their race (Galster, 1990*b*; McIntosh and Smith, 1974; Turner and Mikelsons, 1992). The motivation of the perpetrators of this 'opportunity denying' and 'steering' is attributed as follows 'the primary cause of racial discrimination in housing is that housing agents illegally promote their economic interests by catering to the racial prejudice of their current and potential white customers' (Yinger, 1986, p. 892).

The 'before' and 'after' tests in Britain provide reassurance about the efficacy of legislation in reducing discrimination, on the other hand, significant discrimination was recorded in the United States 25 years after the passage of the Federal Civil Rights Act. The single study of sex discrimination is very small-scale, but the dramatic result calls for replication.

3. Field Experiments in the Product Market

3.1. *The Technique and its History*

Car sales and insurance tests have been conducted using testers making in-person approaches. The methodology of using testers to test for discrimination in the product market is essentially the same as that used for testing in the labour market. Two matched testers approach the car sales or insurance agents seeking information on the sale price of cars or insurance policies, although some tests in Britain have involved sending three matched testers (for example, a Briton, a West Indian and an Hungarian) for insurance inquiries.

The testers differ in one characteristic such as sex, race, or ethnicity. The matched pairs are trained so that both testers in the matched pair make equivalent inquiries to the prospective agents and their background personal characteristics such as income, earnings and credit worthiness are equivalent. They are trained together and carefully supervised throughout the testing period. In making the inquiries there must be some time delay between the testers' approaches, say one hour.

3.2. *Discrimination in Product Markets*

The 1966 PEP study in Britain, sent testers to car insurance providers in various regions of Britain. In the 20 audit tests conducted of car insurance companies, there were no cases of discrimination against the white applicants whereas West Indians were refused insurance cover on 6 occasions and, on 11, were quoted a higher premium (a level of net differential treatment against the West Indian of 85%). Even though the Hungarian tester was quoted a higher premium than the British applicant on 10 of the 20 occasions, the premiums were less than that quoted to West Indians. On average, the premiums quoted were: West Indian £58; Hungarian £49; British £45 (Daniel, 1968, p. 202). Twenty-seven years after the British tests, more extensive telephone tests of home insurance companies, conducted in 1993 in the United States, by the National Fair Housing Alliance body found evidence of similar practices against African–Americans (Yinger, 1998*b*, p. 36).

Ayres, alone, conducted 90 audits of car sales and, together with Siegelman, conducted a further 153 audits (Ayres, 1991; Ayres and Siegelman, 1995). Both series of tests were conducted in Chicago in 1990. A pair of testers, one of whom was always a white male, and the other either a white female, a black male or a black female, were sent to car sales agents. In all, 38 testers were used. The testers were randomly assigned to make the first approach to the car dealer but were not told that they were part of a team. Ayres and Siegelman did not disclose the true nature of the audits to the testers instead they told them that they were involved in testing negotiating behaviour (p. 307). The tests found that white males were quoted significantly lower prices than black male and black female testers and lower, but less significantly so, than white female testers. Ayres and Siegelman report that bargaining reduced the (average) final offer price of a new car, but 'black males were asked to pay $1,100 more than white males, black females $410

more, and white females $92 more' (p. 307). There was no evidence of any effect on these recorded observations from individual tester characteristics (p. 312).

In the case of the provision of insurance the performance of various customer groups does affect the economic outcome, so statistical discrimination is obviously relevant. It is common practice for age to be a determinant of motor vehicle and household insurance premiums for this very reason. Ayres and Siegelman conclude that the disparate treatment of black and white testers may be – '. . .statistical discrimination in which dealers use race and gender as a proxy for the customer's reservation price' (1995, p. 319). Clearly it would be of interest to test the response of dealers to situations involving the white male bargaining over price, and then producing the black male as the purchaser: this is, after all, a scenario which the textbooks would expect.

4. Conclusion

Field experiments of discrimination in the marketplace have extended across 10 countries, several markets and 35 years. They have demonstrated pervasive and enduring discrimination against non-whites and women. Both groups risk being denied employment, housing and insurance purely because of their colour or sex. All countries involved have anti-discrimination legislation dating back to the 1970s, and, in the case of the United States to the 1960s, so clearly there is a need for a serious re-appraisal of equal opportunity policy. For a detailed discussion of policy options see Riach and Rich (2002).

We believe that the foregoing survey has amply demonstrated the significance of carefully-controlled field experiments as a research technique for economists. In the investigation of economic discrimination, field experiments represent an important complement to the conventional regression analysis approach. In the future we expect that field experiments will be applied more widely to age, disability, religion and class. It is appropriate that economists who, after-all, claim pre-eminence in the study of the market, engage in this research and not leave it by default to lawyers and sociologists.

London
Monash University

References

Adam, B. (1981). 'Stigma and employability: discrimination by sex and sexual orientation in the Ontario legal profession', *Canadian Review of Sociology and Anthropology*, vol. 18(2), pp. 216–22.

Akerlof, A. (1980). 'The theory of social custom, of which unemployment may be one consequence', *Quarterly Journal of Economics*, vol. 94, pp. 749–75.

Altonji, J and Blank, R. (1999). 'Race and gender in the labor market', in (O. Ashenfelter and D. Card. eds.), *Handbook of Labor Economics*, vol. 3, chapter 48.

Ayres, I. (1991). 'Fair driving: race and gender discrimination in retail car negotiations', *Harvard Law Review*, vol. 104, pp. 817–72.

Ayres, I. and Siegelman, P. (1995). 'Gender and race discrimination in bargaining for a new car', *American Economic Review*, vol. 85, pp. 304–21.

Becker, G. (1971). *The Economics of Discrimination*. 2nd edn, Chicago: University of Chicago Press.

Bendick Jnr., M. (1996). 'Discrimination against racial/ethnic minorities in access to employment in the United States: empirical findings from situation testing', *International Migration Papers 12*, Geneva, International Labour Office.

Bendick Jnr., M., Brown, L. and Wall, K. (1999). 'No foot in the door: an experimental study of employment discrimination against older workers', *Journal of Aging and Social Policy*, vol. 10, pp. 5–23.
Bendick Jnr., M., Jackson, C. and Reinoso, V. (1994). 'Measuring employment discrimination through controlled experiments', *Review of Black Political Economy*, vol. 23, pp. 25–48.
Bendick Jnr., M., Jackson, C., Reinoso, V. and Hodges, L. (1991). 'Discrimination against Latino job applicants: a controlled experiment', *Human Resource Management*, vol. 30, pp. 469–84.
Blinder, A. (1973). 'Wage discrimination: reduced form and structural estimates', *Journal of Human Resources*, vol. 8, pp. 436–55.
Bovenkerk, F. (1992). *Testing Discrimination in Natural Experiments: a manual for international comparative research on discrimination on the grounds of 'race' and ethnic origin*. Geneva: International Labour Office.
Bovenkerk, F., Gras, M. and Ramsoedh, D. (1995). 'Discrimination against migrant workers and ethnic minorities in access to employment in the Netherlands', *International Migration Papers 4*, Geneva, International Labour Office.
Bovenkerk, F., Kilborne, B., Raveau, F., and Smith, D. (1979). 'Comparative aspects of research on discrimination against non-white citizens in Great Britain, France and the Netherlands', in (J. Berting, F. Geyer and R. Jurkovich, eds), *Problems in International Comparative Research in the Social Sciences*, Oxford: Pergamon Press.
Brown, C. and Gay, P. (1985). *Racial Discrimination 17 Years After the Act*, London: Policy Studies Institute.
Cross, H., Kenney, J., Mell, J., and Zimmermann, W. (1990). *Employer Hiring Practices: Differential Treatment of Hispanic and Anglo Job Seekers*, Washington DC: The Urban Institute Press.
Daniel, W. (1968). *Racial Discrimination in England*, Middlesex: Penguin Books.
Daniel, W. (1970). 'Reply to a 'Note on the testing of discrimination', *Race*, vol. 11, pp. 352–60.
Darity, W. Jnr. and Mason, P. (1998). 'Evidence on discrimination in employment: codes of color, codes of gender', *Journal of Economic Perspectives*, vol. 12, pp. 63–90.
Darity, W. Jnr. and Nembhard, J. (2000). 'Racial and ethnic economic inequality: the international record', *American Economic Review, Papers and Proceedings*, vol. 90 (May), pp. 308–11.
Prada, M. de, Actis, W., Pereda, C. and Perez Molina, R. (1996). 'Labour market discrimination against migrant workers in Spain', *International Migration Papers 9*, Geneva: International Labour Office.
Esmail, A. and Everington, S. (1993). 'Racial discrimination against doctors from ethnic minorities', *British Medical Journal*, vol. 306, pp. 691–2.
Esmail, A. and Everington, S. (1997). 'Asian doctors are still being discriminated against', *British Medical Journal*, vol. 314, p. 1619.
Firth, M. (1981). 'Racial discrimination in the British labour market', *Industrial and Labor Relations Review*, vol. 34, pp. 265–72.
Firth, M. (1982). 'Sex discrimination in job opportunities for women', *Sex Roles*, vol. 34, pp. 265–72.
Fix, M. and Struyk, R. (1993). *Clear and Convincing Evidence: Measurement of Discrimination in America*, Washington DC: The Urban Institute Press.
Fix, M., Galster, G., and Struyk, R. (1993). 'An overview of auditing for discrimination', in (M. Fix and R. Struyk eds.) *Clear and Convincing Evidence: Measurement of Discrimination in America*. Washington DC, The Urban Institute Press.
Fry, E. (1986). *An Equal Chance for Disabled People?* London: The Spastics Society, Campaigns and Parliamentary Department.
Galster, G. (1990*a*). 'Racial discrimination in housing markets during the 1980s: a review of the audit evidence', *Journal of Planning Education and Research*, vol. 9, pp. 165–75.
Galster, G. (1990*b*). 'Racial steering in urban housing markets: a review of the audit evidence', *The Review of Black Political Economy*, vol. 18, pp. 105–29.
Galster, G. and Constantine, P. (1991). 'Discrimination against female-headed households in rental housing: theory and exploratory evidence', *Review of Social Economy*, vol. 69, pp. 76–100.
Goldberg, A., Mourinho, D. and Kulke, U. (1996). 'Labour market discrimination against foreign workers in Germany', *International Migration Papers 7*, Geneva: International Labour Office.
Goldin, C. and Rouse, C. (2000). 'Orchestrating impartiality: the impact of ''blind'' auditions on female musicians', *American Economic Review*, vol. 90, pp. 715–41.
Graham, P., Jordan, A. and Lamb, B. (1990). *An Equal Chance? or No Chance? A study of discrimination against disabled people in the labour market*. London: The Spastics Society.
Gras, M., Bovenkerk, F., Gorter, K., Kruiswijk, P. and Ramsoedh, D. (1996). *Een schijn van kans: Twee empirische onderzoekingen naar discriminatie op grond van handicap en etnische afkomst*. Netherlands: Gouda Quint.
Gunderson, M. (1989). 'Male–female wage differentials and policy responses', *Journal of Economic Literature*, vol. 27, pp. 46–72.
Heckman, J. (1998). 'Detecting discrimination', *Journal of Economic Perspectives*, vol.12, pp. 101–16.
Heckman, J. and Siegelman, P. (1993). 'The Urban Institute audit studies: their methods and findings', in Fix and Struyk (1993).

Henry, F. and Ginzberg, E. (1985). *Who Gets the Work? A test of racial discrimination in employment*. Toronto: The Urban Alliance on Race Relations and the Social Planning Council of Metropolitan Toronto.
Hubbuck, J. and Carter, S. (1980). *Half a Chance? A Report on Job Discrimination against Young Blacks in Nottingham*. London: Commission for Racial Equality.
James, F. and DelCastillo, S. (1992). 'Measuring job discrimination: hopeful evidence from recent audits', *Harvard Journal of African American Public Policy*, vol. 1, pp. 33–53.
Jowell, R. and Prescott-Clarke, P. (1970). 'Racial discrimination and white-collar workers in Britain', *Race*, vol. 11, pp. 397–417.
Kenney, G. and Wissoker, D. (1994). 'An analysis of the correlates of discrimination facing young Hispanic job-seekers', *American Economic Review*, vol. 84, pp. 674–83.
Kim, M. (2002). 'Has the race penalty for black women disappeared in the United States?', *Feminist Economics*, forthcoming August.
La Piere, R. (1934). 'Attitudes vs actions', *Social Forces*, (October), pp. 230–7.
Levinson, R. (1975). 'Sex discrimination and employment practices: an experiment with unconventional job inquiries', *Social Problems*, vol. 22, pp. 533–43.
McIntosh, N. and Smith, D. (1974). *The Extent of Racial Discrimination*, Political and Economic Planning Broadsheet no. 547, London: Political and Economic Planning.
McIntrye, S., Moberg, D., and Posner, B. (1980). 'Preferential treatment in preselection decisions according to sex and race', *Academy of Management Journal*, vol. 23, pp. 738–49.
Mincy, R. (1993). 'The Urban Institute audit studies: their research and policy context', in Fix and Struyk (1993).
Neumark, D., Bank, R. and Van Nort, K. (1996). 'Sex discrimination in restaurant hiring: an audit study', *Quarterly Journal of Economics*, vol. 111, pp. 915–41.
Newman, J. (1978). 'Discrimination in recruitment: an empirical analysis', *Industrial and Labor Relations Review*, vol. 32, pp. 15–23.
Nunes, A. and Seligman, B. (1999). 'Treatment of Caucasian and African-American applicants by San Francisco Bay Area employment agencies: results of a study utilizing ''testers''', *The Testing Project of the Impact Fund*, San Francisco: The Impact Fund.
Nunes, A. and Seligman, B. (2000). 'A study of the treatment of female and male applicants by San Francisco Bay Area auto service shops', *The Testing Project of the Impact Fund*, San Francisco: The Impact Fund.
Oaxaca, R. (1973). 'Male–female wage differentials in urban labor markets', *International Economic Review*, vol. 14, pp. 693–709.
Phelps, E. (1972). 'The statistical theory of racism and sexism', *American Economic Review*, vol. 62, pp. 659–61.
Riach, P. and Rich, J. (1987). 'Testing for sexual discrimination in the labour market', *Australian Economic Papers*, vol. 26, pp. 165–78.
Riach, P. and Rich, J. (1991). 'Testing for racial discrimination in the labour market', *Cambridge Journal of Economics*, vol. 15, pp. 239–56.
Riach, P. and Rich, J. (1991–2). 'Measuring discrimination by direct experimental methods: seeking gunsmoke', *Journal of Post Keynesian Economics*, vol. 14, pp. 143–50.
Riach, P. and Rich, J. (1998). 'Of chicken entrails, anthropology and a realistic social science', *Feminist Economics*, vol. 4, pp. 187–91.
Riach, P. and Rich, J. (2002). 'Women's work or work for women?', in (P. Arestis and S. Dow, eds), *Methodology, Microeconomics and Keynes*, London: Taylor and Francis.
Smeesters, B. and Nayer, A. (1998). 'La discrimination a l'acces a l'emploi en raison de l'origine etrangere: le cas de le Belgique', *International Migration Papers 23*, Geneva: International Labour Office.
Turner, M. and Mikelsons, M. (1992). 'Patterns of racial steering in four metropolitan areas', *Journal of Housing Economics*, vol. 2, pp. 199–234.
Turner, M., Fix, M. and Struyk, R. (1991). *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring*. UI Report 91–9, Washington DC: The Urban Institute Press.
Ward, R. (1969). 'A note on the testing of discrimination', *Race*, vol. 11, pp. 218–33.
Weichselbaumer, D. (2000). 'Is it sex or personality? The impact of sex-stereotypes on discrimination in applicant selection', Working Paper No. 0011, May, University of Linz, Department of Economics.
Wienk, R., Reid, C., Simonson, J. and Eggers, F. (1979). *Measuring Racial Discrimination in American Housing Markets: The Housing Market Practices Survey*, Office of Policy Development and Research, Washington DC: U.S. Department of Housing and Urban Development.
Yinger, J. (1986). 'Measuring discrimination with fair housing audits: caught in the act', *American Economic Review*, vol. 76, pp. 881–93.
Yinger, J. (1998). Testing for Discrimination in Housing and Related Markets. In (M. Fix and M. Turner, eds.), A National Report Card on Discrimination in America. The Role of Testing, Washington D.C., The Urban Institute Press.

Yinger, J. (1991). 'Acts of discrimination: evidence from the 1989 Housing Discrimination Study', *Journal of Housing Economics*, vol. 1, pp. 318–46.
Yinger, J. (1993). 'Access denied, access constrained: results and implications of the 1989 Housing Discrimination Study', in Fix and Struyk (1993).
Yinger, J. (1995). *Closed Doors, Opportunities Lost: The Continuing Costs of Housing Discrimination*, New York: Russell Sage Foundation.
Yinger, J. (1998*a*). 'Evidence on discrimination in consumer markets', *Journal of Economic Perspectives*, vol. 12, pp. 23–40.
Yinger, J. (1998*b*). 'Testing for discrimination in housing and related markets', in (M. Fix and M. Turner, eds), *A National Report Card on Discrimination in America: The Role of Testing*, Washington DC: The Urban Institute Press.
Zimmermann, W. (1993). 'Summary of the Urban Institute's and the University of Colorado's hiring audits', in Fix and Struyk (1993).