APPLIED PSYCHOLOGY

IAAP
THE INTERNATIONAL ASSOCIATION OF APPLIED PSYCHOLOGY

APPLIED PSYCHOLOGY: AN INTERNATIONAL REVIEW, 2012
doi: 10.1111/j.1464-0597.2012.00493.x

Five-Factor Model Personality Measures and Sex-Based Differential Prediction of Performance

Christopher M. Berry* and Anita Kim
Texas A&M University, USA

Ying Wang
University of Western Australia, Australia

Rebecca Thompson
Texas A&M University, USA

William H. Mobley
University of Macau, People's Republic of China

Despite mean differences between sexes, virtually no research has investigated sex-based differential prediction of personality tests in civilian employment samples. The present study investigated the degree to which personality test scores differentially predicted job performance ratings in two managerial samples. In both samples, participants completed a Five-Factor Model personality test and the participants' supervisors, peers, and subordinates provided ratings of participants' task and contextual performance. The current study found sex-based differential prediction in 6.7 per cent of differential prediction analyses in Sample 1, but found no sex-based differential prediction in Sample 2. Across the two samples sex-based differential prediction of performance only occurred 3.3 per cent of the time, which is less than would be expected by chance alone, given alpha = .05. Thus, based on the present study and the extant literature to date, no sex-based differential prediction studies have identified evidence of personality test bias.

INTRODUCTION

Personality testing is widespread in organisations. For instance, personality testing is a $400 million industry, and at least 30 per cent of all US organi-

* Address for correspondence: Christopher M. Berry, Texas A&M University, Department of Psychology, 4235 TAMU, College Station, TX 77843, USA. Email: cmberry@tamu.edu

We thank Ronald C. Page of Assessment Associates International and William H. Mobley of Mobley Group Pacific for providing the data for this project. We also thank Paul Sackett and Stephanie Payne for thoughtful comments on an earlier version of this manuscript.

sations use personality tests for hiring or related practices (Paul, 2004). Given this widespread use, it is worth asking whether personality testing is a fair personnel selection practice. Since Barrick and Mount's (1991) seminal meta-analysis of the criterion-related validity of the Five-Factor Model (FFM) of personality, much research has focused on personality tests as personnel selection tools (e.g. Costa, 1996; Saville, Sik, Nyfield, Hackston, & MacIver, 1996). Many studies are devoted to issues that have bearing on the fair use of personality tests in personnel selection; issues such as criterion-related validity (Barrick, Mount, & Judge, 2001), faking and social desirability (Ones & Viswesvaran, 1998), and race and sex differences in mean personality test scores (Ash, Baehr, Joy, & Orban, 1988; Bobko, Roth, & Potosky, 1999; Feingold, 1994; Hough, Oswald, & Ployhart, 2001). Of particular interest for the present study are meta-analytic results outlining mean differences between sexes in personality tests scores for numerous traits (Feingold, 1994; Hough et al., 2001). Although sex differences in predictor scores do not in and of themselves signal predictive bias (also termed "differential prediction"), such sex differences make predictive bias a more salient concern (Saad & Sackett, 2002). The present study investigates whether FFM personality test scores differentially predict job performance for men versus women, and what might cause this sex-based differential prediction.

Differential Prediction

Differential prediction occurs when the regression lines relating a test score to a relevant criterion are not equivalent for subgroups (Society for Industrial and Organizational Psychology [SIOP], 2003), meaning that the test scores do not predict the criterion the same for each subgroup. If such predictive bias existed, it could affect conclusions regarding whether the use of personality testing in personnel selection is fair or even perhaps legal (e.g. if test scores do not predict job performance as well for a minority subgroup, the use of test scores for personnel selection might be deemed unfair). Given the amount of research devoted to differential prediction of cognitive ability tests (Sackett, Borneman, & Connelly, 2008), it is surprising that there exists only one published study (Saad & Sackett, 2002) related to sex-based differential prediction of personality tests.

Differential prediction is typically assessed using a regression model wherein the within-group regression lines relating test scores to job performance criterion scores are compared in terms of their relative slopes and intercepts. Differences between the two subgroups' (i.e. sexes') regression slopes and/or intercepts indicate predictive bias (SIOP, 2003). If the regression slopes differ between sexes, this would mean that the test score is not as strongly related to the criterion for the sex with the smaller regression slope, and thus inferences drawn from test scores might not be as accurate for that

sex. If only the regression intercepts differ, this suggests that the sex with the higher intercept's criterion scores would be under-predicted by the overall sample common regression line, and thus the use of a common regression line (which is mandated by the 1991 Civil Rights Act) would be a disadvantage for that sex. Because women are the traditionally disadvantaged group compared to men in the workplace (e.g. Sackett & Wilk, 1994), of most concern from an applied personnel selection perspective is if women's performance is under-predicted by the common regression line or if the regression slope is less steep for women. Therefore, the present study focuses on sex-based differential prediction in the form of weaker slopes or under-prediction of job performance for women, although we recognise that the opposite (weaker slopes or under-prediction of performance for men) is also possible.

Should Organisations Be Concerned about Sex-based Differential Prediction of Personality Tests?

Differential prediction is generally of greatest concern for predictors exhibiting criterion-related validity, as these are the predictors that are most likely to be used in personnel selection systems, and for predictors exhibiting mean differences between subgroups, as these are the predictors most likely to result in adverse impact. For example, race-based differential prediction has been such a concern for cognitive ability tests because cognitive ability tests are highly valid personnel selection tools on which certain racial/ethnic subgroups score much lower. FFM personality tests also fit both of these criteria with regard to sex differences. First, most of the FFM traits are related to at least one of the main dimensions of job performance. For instance, conscientiousness and emotional stability predict task performance (Barrick et al., 2001). Agreeableness predicts organisational citizenship behaviors (Hurtz & Donovan, 2000) and counterproductive work behaviors (Berry, Ones, & Sackett, 2007a). Although openness and extraversion do not predict task or counterproductive performance, they do predict leadership and managerial performance, respectively (Barrick et al., 2001; Judge, Bono, Ilies, & Gerhardt, 2002), and openness predicts citizenship performance (Chiaburu, Oh, Berry, Li, & Gardner, 2011). In addition, Hogan and Holland (2003) demonstrated that when FFM personality traits and performance criteria are aligned using socioanalytic theory, all of the FFM traits predicted relevant criteria with true validities exceeding .30. Thus, each of the FFM traits is viable in at least some personnel selection settings.

FFM personality tests also fit the second criterion regarding subgroup differences. Meta-analyses by Feingold (1994) and Hough et al. (2001) outline a number of traits, at both the FFM and facet levels, on which men have higher mean scores than women. Even for traits on which women score higher than men, sex-based differential prediction could occur if there are not

comparable mean differences in favor of women on criterion performance or if the trait scores do not predict performance as strongly for women. Regardless, the main point is that the FFM traits are viable predictors of job performance that exhibit mean sex differences. Therefore, the general lack of research on sex-based differential prediction of FFM personality tests is an oversight, and the present study investigates whether such sex-based differential prediction exists in two employee samples.

Existing Evidence

There exists almost no research on sex-based differential prediction of personality tests. There are a number of possible reasons for this dearth. One possibility is that differential prediction analyses originally gained popularity as a method for assessing predictive bias of cognitive ability tests for African American and White subgroups (Cleary, 1968). Thus, when researchers think of predictive bias and differential prediction, they may think of race and cognitive ability, not sex and personality. Another possibility is that personality tests were not originally developed for making predictions about job applicants in employment settings and only relatively recently gained popularity as a personnel selection tool. So, considerations of predictive bias, a concept mostly confined to educational and employment testing, was not always particularly relevant for personality tests. However, given the increasing use of personality tests in personnel selection and employee development, the lack of research on sex-based differential prediction must be remedied.

To our knowledge, only one study has investigated sex-based differential prediction of personality tests. Saad and Sackett (2002) reported sex-based differential prediction analyses separately for nine military jobs. Saad and Sackett included three different personality traits from the Army's Assessment of Background and Life Experiences instrument (Peterson, Hough, Dunnette, Rosse, Houston, Toquam, & Wing, 1990): Adjustment (similar to the FFM trait of Emotional Stability); Dependability (a facet of the FFM trait of Conscientiousness); and Achievement-Orientation (another facet of the FFM trait of Conscientiousness); and five different soldier performance criteria. In all, Saad and Sackett carried out 135 sex-based differential prediction analyses (9 jobs × 3 traits × 5 criteria). On the one hand, *slope differences* between sexes were not found at greater than chance levels (i.e. significant slope differences would be expected 5% of the time simply due to chance, based on alpha equal to .05). On the other hand, *intercept differences* between sexes were found in about one-third of the analyses. These intercept differences were almost always in the form of women having a lower intercept (meaning the common regression line over-predicted female performance). Further, almost all of the cases of sex-based differential prediction were concentrated within the Effort and Leadership criterion, which "reflects the

level of individual effort exerted over all job tasks, perseverance under adverse conditions, leadership qualities, and support of peers" (Saad & Sackett, 2002, p. 669). This suggests some issue with that criterion rather than bias in the personality tests. In all, Saad and Sackett did not find consistent evidence compatible with personality test bias, and any bias they did find actually resulted in over-prediction, rather than the more concerning under-prediction, of female performance.

Although Saad and Sackett (2002) stands as an excellent example of the kind of sex-based differential prediction research that must be carried out, there remain a number of unanswered research questions. One set of unanswered questions pertains to the personality traits studied in Saad and Sackett. In particular, Saad and Sackett used a limited set of personality traits. Their personality measures were only related to two of the FFM model traits (Conscientiousness and Emotional Stability). Thus, the present study investigates sex-based differential prediction of all of the FFM traits instead of just a subset. Further, two of Saad and Sackett's personality measures (Dependability and Achievement-Orientation) were facet-level personality measures. It has been documented that relationships at the facet level of personality measurement do not always translate to the FFM level (e.g. Ones & Viswesvaran, 1996). Therefore, the present study focuses on the more common FFM framework, which is another important advancement beyond Saad and Sackett.

Similarly, Saad and Sackett used a limited set of performance criterion measures. Their criterion measures were all based on supervisor ratings, work samples, and personnel files. Although these are clearly important criterion measures, the degree to which their results extend beyond those criterion sources remains unanswered. The present study focuses on peer and subordinate ratings of performance in addition to supervisor ratings. Although supervisor ratings are more commonly used for administrative purposes within organisations than peer or subordinate ratings, this is not always the case. For instance, in addition to their common use for employee development, 360 degree performance ratings are increasingly being used for administrative purposes (London & Smither, 1995). In addition, Oh and Berry (2009) presented evidence that supervisor ratings represent a deficient performance criterion and that adding peer and subordinate ratings to supervisor ratings helps address this criterion deficiency issue. Thus, peer and subordinate ratings can be seen as viable performance criteria and therefore the present study's focus on these rating sources, in addition to supervisor ratings, provides a broader picture of personality tests' sex-based differential prediction of job performance.

Also, Saad and Sackett used a military sample. The military represents a relatively unique context in that it is strongly male-dominated and hierarchical in nature. So, the degree to which Saad and Sackett's results generalise to

civilian employment settings is an open question. Therefore, the present study's focus on civilian employee samples represents another advancement of the present study over previous research. More fundamentally, one study cannot provide a definitive answer regarding sex-based differential prediction of personality tests. Relatedly, Saad and Sackett's analyses were carried out within jobs, and female sample sizes ranged from 31 to 281 (mean $N = 89$; samples composed of between 6.6 and 54.7% women). The fact that Saad and Sackett is only a single study with a moderate sample highlights the need for future incremental research, and the present study addresses this need by carrying out sex-based differential prediction analyses in two new samples.

An Agency–Communion Framework for Understanding Sex-based Differential Prediction

The above section highlighted the need for continued sex-based differential prediction research. As new studies are carried out, it would be useful to have a framework for predicting what personality traits and performance criteria are most likely to be associated with sex-based differential prediction. Therefore, the present study introduces the agency–communion framework for understanding sex-based differential prediction, if it exists. Sex-based differential prediction can be a function of some form of bias in either the predictor or the criterion. We posit that predictor and/or criterion bias against women is more likely to manifest when the predictor (i.e. personality trait) or criterion (i.e. job performance) reflects content that is counterstereotypic of women.

The growing body of research in sexism and gender stereotyping reveals that men are regarded as *agentic*, whereas women are regarded as *communal* (Burgess & Borgida, 1999). In other words, men are regarded as having more forceful, instrumental qualities like being competitive and being assertive, and women are seen as being more passive, warm, and concerned about others' welfare. Therefore, agentic characteristics are counterstereotypic of women. Predictors or criteria with strong agentic content may exhibit different measurement properties for men and women. For instance, Sheppard, Han, Colarelli, Dai, and King (2006) found that, holding personality trait scores constant, men were more likely to endorse personality items with agentic content (e.g. competitiveness, assertiveness), and that this accounted for sex-based differential item functioning in the Hogan Personality Inventory (Hogan & Hogan, 1992). Similarly, when women exhibit counterstereotypic behaviors such that they behave agentically, there is a resulting backlash; they are penalised with respect to competence and likability ratings (Rudman, 1998) and are even sabotaged on subsequent tasks (Rudman & Fairchild, 2004). Research specifically focused on job performance evaluation has also supported this idea that women are often penalised when there

is a lack of fit between stereotypes of women and job roles (e.g. Davison & Burke, 2000; Eagly, Makhijani, & Klonsky, 1992; Lyness & Heilman, 2006). If personality test scores or job performance criteria have different psychological meaning for men versus women as a function of agentic content, then this could affect the degree to which the personality trait scores predict job performance Therefore, we suggest that sex-based differential prediction of job performance is more likely when the personality traits and/or job performance criteria contain agentic content.

Present Study

The current study carried out an investigation of sex-based differential prediction using an FFM personality test in two civilian employment samples, and included performance ratings from three sources: supervisors, peers, and subordinates. Based on Saad and Sackett's (2002) null test bias results, it would not be surprising to find null test bias results in the current study as well. However, if evidence of sex-based differential prediction is found, and it is a result of either test or criterion bias, we suggest that the agency–communion framework is a useful tool for hypothesising which traits and criteria are most likely to be associated with sex-based differential prediction. That is, we posit that sex-based differential prediction is more likely to be found for certain FFM traits and performance criteria, based on the degree to which they are relatively agentic traits/criteria. For instance, regarding bias as a function of personality tests, sex-based test bias might be a function of the contaminating effects of agentic content in personality test items (Sheppard et al., 2006), and therefore sex-based differential prediction would be expected to be more common for relatively agentic traits. Berry, Page, and Sackett (2007b), based in part on arguments put forth by Paulhus and John (1998), made the case that Extraversion, Openness to Experience, and Emotional Stability are relatively agentic traits. Therefore, if sex-based differential prediction exists in the present study, we hypothesise:

> *Hypothesis 1*: Sex-based differential prediction will occur more often for relatively agentic traits (Extraversion, Openness, or Emotional Stability) than for relatively communal traits (Agreeableness or Conscientiousness).

Regarding criterion bias, the agency–communion framework suggests that performance ratings of women are most likely to be contaminated by sex bias when those ratings regard relatively agentic dimensions of performance that are counterstereotypic of women. The job performance domain is often conceptualised as comprising task performance (i.e. behaviors focused on structuring work and getting things done) and contextual performance (i.e. behaviors focused on facilitating the psychological and social contexts of work and getting along with others) dimensions (e.g. Borman & Motowidlo,

1997; Johnson, 2001; Motowidlo, Borman, & Schmit, 1997; Motowidlo & Van Scotter, 1994; Oh & Berry, 2009; Van Scotter & Motowidlo, 1996). Oh and Berry (2009) made the case that task performance dimensions are reflective of a motivation to achieve status (i.e. agency) while contextual performance dimensions are reflective of a motivation to get along with others (i.e. communion). Therefore, if sex-based differential prediction exists in the present study, we hypothesise:

> *Hypothesis 2*: Sex-based differential prediction will occur more often for relatively agentic performance dimensions (task performance) than for relatively communal performance dimensions (contextual performance).

Adding an additional layer of complexity to these predictions, the current study included performance ratings from three sources: supervisors, peers, and subordinates. It is possible that ratings from supervisors, peers, and subordinates are differentially influenced by their own motives and biases with respect to the ratee. Based on socioanalytic theory, Oh and Berry (2009) made the case that a performance rater's judgments are based, in part, on the degree to which the ratee meets the rater's expectations and promotes the rater's agenda. For example, in the case of ratings of managerial performance, as in the current study, the agendas of supervisor raters likely reflect a need for production from the managers beneath them and therefore are more likely to be weighted toward "getting ahead" behaviors (i.e. agentic behaviors) of managers. On the other hand, Oh and Berry stated that peer and subordinate raters' agendas likely reflect more of a desire for considerate managers who are easy to get along with (given the benefits of such considerate manager behavior to peers and subordinates), and therefore are more likely to be weighted toward "getting along" behaviors (i.e. communal behaviors) of managers. Sex-based differential prediction is most likely when the performance ratings are weighted toward agentic behaviors that are counterstereotypic of women. Thus, to the degree that sex-based differential prediction exists in the present study, we hypothesise:

> *Hypothesis 3*: Sex-based differential prediction will occur more often when the relationship in question involves supervisor ratings.

METHOD

Participants

Sample 1. Two hundred and seventy-seven managers at a large US energy company comprised Sample 1. These were "middle-managers" occupying a diverse group of positions at organisational levels above front-line supervisor, but below the level of vice president. Some were plant general

managers, while others held managerial positions in a wide range of departments such as human resources, information technology, finance, and public affairs and regulatory services.

The managers participated as part of a leadership development program in which participants completed an FFM-based inventory and their performance levels were assessed on a customised 360 degree performance rating system. Of the 277 managers, 264 provided their sex (60 female), and thus comprised the final sample. One hundred and fifty participants provided their age (mean = 50.7, SD = 7.8).

Findings from the data in Sample 1 have been previously published in two manuscripts. Berry et al. (2007b) focused on the moderating role that Self-deceptive Enhancement and Impression Management play in the relationship between Overall Job Performance and the FFM personality variables (Emotional Stability, Extraversion, Openness, Agreeableness, and Conscientiousness). Oh and Berry (2009) focused on the potential of multisource job performance ratings to enhance estimates of the relationships between FFM personality variables and Task and Contextual Performance. The current manuscript focuses on the potential for sex-based differential prediction in personality–performance relationships, an issue that was not addressed in Berry et al. (2007b) or Oh and Berry (2009). A number of things in the current study mitigate the potential issues associated with using previously published data. For one, there might be concern that the results of Berry et al. (2007b) or Oh and Berry (2009) confound results of the present study. This is unlikely the case. Berry et al. (2007b) focused on the effects of response distortion on personality validity. If response distortion were confounded with sex (i.e. affected personality validity more for one sex than the other), this would be expected to exacerbate validity/prediction differences between sexes, not suppress them. That no sex-based differential prediction was found in the current study argues against a confounding with response distortion. Oh and Berry (2009) demonstrated that single-source supervisor ratings of performance are potentially deficient criterion measures, but the current study did not rely on single-source supervisor ratings. A second concern might be that there is something idiosyncratic about this dataset and thus the continued use of it is misleading. This does not appear to be the case as the same pattern of a lack of sex-based differential prediction was repeated in a large, independent sample (Sample 2) in the present study.

Sample 2. Seven hundred and fifty-four Chinese executive MBA (EMBA) students from a diverse set of organisations and management levels comprised Sample 2. As part of a developmental class assignment, the EMBA students participated in a 360 degree performance rating exercise. First, the EMBA students rated their own performance on an online performance rating form. Then, each EMBA student enlisted their supervisors,

peers, and subordinates at the EMBA student's place of employment to rate the EMBA student's performance using the same online form. The executive MBA students then were invited to complete an FFM-based personality inventory during their 360 degree feedback session. Of the 754 participants, 134 were female. Six hundred and ninety-nine participants provided their age (mean = 38.5, *SD* = 4.5).

Measures

Sample 1. The Work Behavior Inventory (WBI; Page, 2009) 1.0 was used to measure participants' FFM personality traits using a 5-point Likert scale. The WBI 1.0 is a 240-item occupational-purpose personality instrument that has been used in past organisational research (e.g. Berry et al., 2007b; Oh & Berry, 2009). The WBI 1.0 comprises 20 facet-level scales, 18 of which map onto the FFM traits (see the Appendix for example items). Scale scores were calculated for participants on each of the FFM traits: Emotional Stability, Extraversion, Openness to Experience, Conscientiousness, and Agreeableness. The company that owns the WBI 1.0 and made the data available for the present study converted scale scores to a *T*-score metric (where the mean and standard deviation are 50 and 10, respectively, based on test norms), and thus the FFM scale scores in the present study are in such a metric. Alpha reliabilities for the FFM scales ranged from .92 to .95.

The job performance rating form was a proprietary instrument containing 20 separate items, each representing a job performance dimension (see Table 1 for definitions of each of the 20 performance dimensions). Raters were provided with construct descriptions of each of the performance dimensions, along with examples of less effective versus highly effective behaviors. Raters were then asked to rate participants on each of the 20 dimensions on a 9-point scale (1 = needs substantial development, 9 = extremely effective). Participants' job performance was rated by an average of 12.44 raters (one supervisor, 6.75 peers, and 4.69 subordinates). Supervisors, peers, and subordinates used the same rating form. Each participant was rated by only one supervisor, and only average peer ratings and average subordinate ratings were made available to the researchers; so interrater agreement statistics could not be calculated.

Exploratory factor analyses (EFAs) using promax rotation were conducted in order to determine whether the 20 performance dimensions would map onto task and contextual performance factors (see Table 2). Separate EFAs were carried out for each of the three ratings sources: supervisor, peer, and subordinate ratings. For each of the three ratings sources, the same three interpretable factors (accounting for between 68.0 and 72.6% of the variance) emerged. Performance dimensions 1–6 (which all resembled task performance in that they reflect core job tasks of managers and are focused on

TABLE 1
Definitions of Each of the Performance Dimensions in Samples 1 and 2

Performance dimensions	*Definitions*
Sample 1	
1. Continuous Learning	Comprehends quickly; applies new concepts; actively works to continuously gain understanding and knowledge to improve approaches, technologies, solutions, market conditions and customer needs.
2. Technical Orientation	Demonstrates expertise in own technical field; serves as a technical resource for others.
3. Problem Analysis	Analyzes and solves problems using a sound problem-solving strategy: gathers information, analyzes the problem using critical thinking, generates and selects solutions, checks results.
4. Strategic Planning	Leverages strategy and objectives to drive goals and plans; sets clearly defined objectives; produces timely, comprehensive project plans with action steps. Plans for future problems and opportunities by forecasting business trends and outside forces; considers benefits of several options by using resources and focusing efforts on critical components.
5. Business Acumen	Gets work done efficiently through formal and informal channels; displays broad understanding of business practices and policy.
6. Leading, Modeling, & Vision	Provides leadership by example; defines a vision and engages others to implement the vision; sets a strong leadership role by walking the talk.
7. Team Building	Supports team efforts; builds a spirit of participation and belonging; builds group cohesiveness by emphasising team objectives and reinforcing cooperation.
8. Coaching & Mentoring	Provides challenging assignments with clear and constructive feedback to employees to yield high performance; acts as a positive mentor; fosters development in others with an understanding of individual's career aspirations; brings out the best in individuals regardless of differences in background or experiences.
9. Relationship Partnering	Builds and maintains effective working relationships; has a wide and effective network; quickly establishes rapport with others.
10. Delegation	Plans and assigns work and work responsibility to direct reports in order to balance their development and productivity; provides resources (human and capital) to help subordinates succeed in delegated responsibilities.
11. Approachability	Is approachable; actively listens to the issues and concerns of others; manages by "walking around" to collect information and make responsive decisions; provides an environment that allows others to be comfortable talking about sensitive issues.
12. Leveraging Conflict	Openly manages conflict and disagreement through collaborative discussion to reach positive conclusions or decisions; arrives at constructive solutions while maintaining positive working relationships; seeks win-win situations.
13. Negotiating Resolution	Diplomatically explores common and opposing options to reach mutually acceptable positive solutions; is persuasive and clear in addressing negotiable items.

TABLE 1
Continued

Performance dimensions	*Definitions*
14. Initiative & Risk Taking	Demonstrates a high motivation for succeeding; shows effort and drive in the face of obstacles; sets aggressive goals for oneself and works hard in achieving those goals; is internally driven; believes calculated risk taking is necessary in a competitive marketplace; a sense of urgency prevails.
15. Results Driven	Drives for successful results; moves tasks and assignments toward closure.
16. Communication	Whether speaking or writing, clearly articulates the key points of an issue; interpersonally holds the attention of others; takes recipient's needs into account; addresses and listens to others in a respectful manner.
17. Process Management	Understands how to design efficient work flow; identifies opportunities for synergy and integration; attains higher productivity yields with fewer resources and simpler processes.
18. Customer Focus	Responsively addresses the needs of the customer; accurately diagnoses customer needs; communicates effectively with customers; establishes customer rapport.
19. Decision Making	Makes quality decisions based on a mixture of analysis, wisdom, judgment, and due diligence.
20. Adaptability	Appropriately changes one's strategy in response to new information; continuously adapts to changes; deals with uncertainty and vagueness; decides and acts without having the picture totally defined; is receptive to understanding cultural and individual differences.
Sample 2	
1. Strategic Thinking	Maintaining a big-picture view of the industry and organisation; quick to recognise trends and changes; envisioning and following clear goals.
2. Analysis and Problem Solving	Analyzing situations, identifying alternative solutions, and developing specific items; capable of making high-quality decisions.
3. Organising and Planning	Assigning responsibilities and coordinating tasks; providing direction and scheduling activities.
4. Executing and Implementing	Tackling problems head-on and managing change; integrating efforts across functions and assigning clear authority and accountability.
5. Customer Focus	Ensuring that team members stay alert to external customers' expectation; developing useful contacts with a range of people in a variety of locations.
6. Self-Management	Maintaining personal ethical standards and directing oneself in one's work and career development.
7. People Management	Teaming with and developing others in the organisation.
8. Innovation & Change	Feeling comfortable in fast-changing environments, being willing to take risks and to consider new and untested approaches.
9. Influence & Communication	Getting the message through clearly and in ways that lead others to share their perspective and reach agreement, being attentive and responsive to other people's feelings.
10. Interpersonal Relations	Staying alert to others' needs and concerns; maintaining good interpersonal relations with others.

TABLE 2
Results of the Exploratory Factor Analyses of Performance Dimensions in Samples 1 and 2

	Supervisor ratings			*Peer ratings*			*Subordinate ratings*		
	F1	*F2*	*F3*	*F1*	*F2*	*F3*	*F1*	*F2*	*F3*
Sample 1									
Eigenvalues	11.4	1.2	.98	11.4	1.6	1.2	12.3	1.4	.86
Percent of variance	57.2	5.9	4.9	57.2	7.9	5.9	61.4	6.9	4.3
Factor loadings									
1. Continuous Learning	.507	−.080	.396	.756	−.094	.117	.762	.047	.048
2. Technical Orientation	.685	−.060	.094	.859	−.095	−.115	.810	−.104	−.006
3. Problem Analysis	.914	.201	−.267	.742	.242	−.192	.797	.149	−.148
4. Strategic Planning	.528	−.124	.406	.763	−.072	.191	.861	−.087	.056
5. Business Acumen	.468	.010	.325	.640	.048	.233	.782	−.111	.164
6. Leading, Modeling, & Vision	.416	.275	.259	.520	.117	.321	.592	.286	.077
7. Team Building	−.150	.792	.152	−.133	.854	.142	−.049	.941	−.016
8. Coaching/Mentoring	.182	.652	.030	.255	.616	−.072	.167	.803	−.107
9. Relationship Partnering	−.086	.620	.314	−.178	.940	.111	−.020	.861	.075
10. Delegation	.139	.650	.081	.431	.599	−.268	.463	.536	−.214
11. Approachability	.101	.699	−.069	−.123	.840	.126	−.219	.944	.066
12. Leveraging Conflict	−.072	.285	.457	−.115	.038	.855	−.015	−.043	.933
13. Negotiating Resolution	.047	.304	.473	.090	.183	.609	.237	.234	.460
14. Initiative & Risk Taking	.229	−.092	.756	.849	−.243	.085	.880	−.147	.042
15. Results Driven	.219	.069	.594	.767	.016	.062	.694	.080	.051
16. Communication	.085	.149	.624	.369	.081	.468	.314	.320	.229
17. Process Management	.228	.548	.116	.613	.384	−.130	.624	.107	.129
18. Customer Focus	−.249	.317	.670	.118	.508	.204	.051	.382	.395
19. Decision Making	.368	.424	.039	.796	.131	−.095	.803	.157	−.143
20. Adaptability	.137	.053	.669	.559	.225	.101	.342	.351	.183

TABLE 2
Continued

	Supervisor ratings			*Peer ratings*			*Subordinate ratings*		
	F1	*F2*	*F3*	*F1*	*F2*	*F3*	*F1*	*F2*	*F3*
Sample 2									
Eigenvalues	8.1	.53	–	8.5	.45	–	8.5	.46	–
Percent of variance	80.8	5.3	–	84.5	4.5	–	84.5	4.6	–
Factor loadings									
1. Strategic Thinking	.820	.093	–	.896	.041	–	.917	.001	–
2. Analysis & Problem Solving	.852	.093	–	.850	.121	–	.865	.117	–
3. Organising & Planning	.950	−.031	–	.940	.010	–	.895	.061	–
4. Executing & Implementing	.945	−.066	–	.918	−.005	–	.924	−.033	–
5. Customer Focus	.702	.176	–	.596	.308	–	.697	.206	–
6. Self Management	.564	.397	–	.524	.461	–	.568	.421	–
7. People Management	.572	.394	–	.541	.438	–	.534	.450	–
8. Innovation & Change	.717	.245	–	.704	.278	–	.743	.225	–
9. Influence & Communication	.401	.570	–	.422	.554	–	.430	.554	–
10. Interpersonal Relationships	−.106	.965	–	−.087	.982	–	−.094	1.01	–

Note: F1 = Factor 1 (i.e. task performance), F2 = Factor 2 (i.e. contextual performance), F3 = Factor 3 (i.e. "third factor"); dashes represent "not applicable" (i.e. the third factor did not emerge in Sample 2).

structuring work and getting things done; Oh & Berry, 2009) loaded most strongly on the first factor for all three ratings sources. Therefore, task performance composite variables were created for each rating source by summing ratings on performance dimensions 1–6. Performance dimensions 7–11 (which all resembled contextual performance in that they focused on facilitating the social and psychological contexts of work, interpersonal relations, and relationship building; Oh & Berry, 2009) loaded most strongly on the second factor for all three rating sources. Therefore, contextual performance composite variables were created for each rating source by summing ratings on performance dimensions 7–11. Performance dimensions 12 and 13 (which dealt with leveraging conflict and negotiating resolution) loaded most strongly on the third factor for all rating sources, so "third-factor" composite variables were created for each rating source by summing ratings on performance dimensions 12 and 13. Because this third factor did not fit our theoretical task and contextual performance model of job performance ratings, we do not report full analyses for this third factor. Although we do not report them, we did run full analyses (e.g. sex-based differential prediction analyses) for the third factor composites and the exact same pattern of results emerged as with the task and contextual performance factors, so study conclusions remain the same with or without this third factor. Full sex-based differential prediction analysis results for this third factor are available upon request from the first author.[1] Performance dimensions 14–20 loaded most strongly on different factors depending on the rating source and therefore cannot definitively be called task versus contextual performance dimensions. Thus, dimensions 14–20 were excluded from further analyses. In all, six performance criterion composite variables were created and used in sex-based differential prediction analyses: task and contextual performance composite variables for supervisor, peer, and subordinate ratings.

Intercorrelations between supervisor, peer, and subordinate ratings of task performance ranged from .16 to .42; and intercorrelations between supervi-

[1] Another possible way to form task and contextual performance factors would have been to only extract the first two (task and contextual performance) factors in the EFAs. In order to investigate what effects this would have on study conclusions, we carried out the EFAs in this fashion. Because there was not a third factor to load on, some of the performance dimensions that had loaded most strongly on the third factor instead loaded most strongly on the task and contextual performance factors across ratings sources. We created task and contextual performance composite variables including the new dimensions that now loaded most strongly on the task and contextual performance factors and carried out sex-based differential prediction analyses using these new composites as criteria. The pattern of results remained exactly the same as when we allowed there to be a third factor and excluded performance dimensions loading on the third factor. Full results are available on request from the first author. The main point is that, regardless of how the task and contextual performance factors were formed, sex-based differential prediction analysis results remained the same.

sor, peer, and subordinate ratings of contextual performance ranged from .16 to .31 (see Table 3). These low-to-moderate intercorrelations are very similar to those found in Conway and Huffcutt's (1997) meta-analysis of these same relationships. These low-to-moderate intercorrelations suggest that the ratings from the three sources were relatively independent of each other, likely due to differences in rater agendas and/or opportunity to observe performance behaviors of ratees (Oh & Berry, 2009). Therefore, the three rating sources were considered separately in the present study.

Sample 2. The WBI 2.0, Chinese-Language Version (Page, 2009) was used to measure participants' FFM personality traits using a 5-point Likert scale. Scale scores were calculated for participants on each of the FFM traits: Emotional Stability, Extraversion, Openness to Experience, Conscientiousness, and Agreeableness. As in Sample 1, FFM scale scores are in a *T*-score metric. Alpha reliabilities for the FFM scales ranged from .84 to .91.

The WBI 2.0 is an updated version of the WBI 1.0. WBI 2.0 differs from WBI 1.0 in that some item content is different, and three facet-level scales (i.e. sub-scales for the FFM traits) were added to the inventory (see Page, 2009). The WBI 2.0, Chinese Language Version is a Chinese translation of the WBI 2.0, English Language Version. A number of studies have documented the comparability of the Chinese and English versions of the WBI. Liang and Yang (2006) found the scale reliabilities to be essentially comparable across the two versions, with the average reliability being .86 for the English version and .81 for the Chinese version. Bilingual persons taking both the English and Chinese versions yielded high same-scale consistency (mean $r = .92$) across the two language forms (Thompson, Hartmann, Vang, & Tubré, 2008). In addition, identical FFM structure has been replicated across the English and Chinese versions (Page, 2009).

Participants' job performance was rated by an average of 10.96 raters (1.10 supervisors, 3.45 peers, and 5.65 subordinates). Average supervisor, average peer, and average subordinate ratings were used in the present study. The performance rating form was developed as a generic 360 degree performance assessment that is suitable for managerial populations working in the Chinese context (Wang, Fang, & Mobley, 2006). This performance rating form was different from that used in Sample 1. The performance rating form contained 10 separate job performance dimensions (see Table 1 for definitions of each of the 10 performance dimensions). Each of the 10 performance dimensions consisted of 5 to 15 behavioral items. Supervisors, peers, and subordinates used the same rating form. Raters were required to rate how frequently the ratee demonstrated each behavior at work as described in the item, using a 5-point scale (1 = never, 5 = always). Raters' ratings on specific behavioral items were then averaged to generate dimensional scores.

TABLE 3
Correlations between All Variables in Samples 1 and 2

	Sex	*EX*	*AG*	*OP*	*CO*	*ES*	*Sup TP*	*Sup CP*	*Peer TP*	*Peer CP*	*Sub TP*	*Sub CP*
Sex		.09	.03	.14*	.02	.19*	−.06	−.08	.00	−.09	−.05	−.06
Extraversion	−.07	1.00	.46*	.60*	.56*	.55*	.19*	.18*	.24*	.14*	.05	.03
Agreeableness	−.10*	.42*	1.00	.29*	.48*	.58*	.01	.15*	.03	.18*	.04	.16*
Openness	.20*	.57*	.28*	1.00	.44*	.55*	.20*	.09	.32*	.09	.09	−.10
Conscientiousness	−.09	.52*	.51*	.42*	1.00	.50*	.09	.15*	.20*	.08	.04	.06
Emotional Stability	−.04	.39*	.54*	.42*	.47*	1.00	.15*	.17*	.21*	.18*	.02	.02
Supervisor Ratings: Task Performance	−.01	.01	.05	.04	.05	.07	1.00	.73*	.42*	.18*	.16*	.04
Supervisor Ratings: Contextual Performance	−.02	.05	.16*	.05	.07	.21*	.81*	1.00	.32*	.31*	.13*	.16*
Peer Ratings: Task Performance	−.01	.09*	.09*	.05	.11*	.05	.28*	.32*	1.00	.70*	.27*	.13
Peer Ratings: Contextual Performance	.02	.07	.19*	.03	.04	.12*	.21*	.38*	.85*	1.00	.14*	.24*
Subordinate Ratings: Task Performance	−.00	.04	.06	.01	.08*	.09*	.18*	.13*	.27*	.14*	1.00	.73*
Subordinate Ratings: Contextual Performance	.08	.03	.18*	−.02	.04	.16*	.14*	.28*	.23*	.35*	.85*	1.00

Note: Sample 1 and 2 results above and below the diagonal, respectively; for "Sex", 1 = Male, 0 = Female.

* Significant at $p < .05$.

Exploratory factor analyses (EFAs) using promax rotation were conducted on the 10 performance dimensions in the same manner as in Sample 1 (see Table 2). Across all three rating sources, the same two interpretable factors (accounting for between 86.1 and 89.1% of the variance) emerged, suggesting that two-factor solutions accounted for most of the variance in performance ratings (an interpretable third factor did not emerge, as in Sample 1). Additionally, for supervisors, peers, and subordinates, the same sets of performance dimensions hung together. For all three rating sources, performance dimensions 1–8 (which all resembled task performance; Oh & Berry, 2009) loaded most strongly on the first factor. Therefore, task performance composite variables were created for each rating source by summing ratings on performance dimensions 1–8. Performance dimensions 9–10 (which each resembled contextual performance; Oh & Berry, 2009) always loaded most strongly on the second factor, and therefore contextual performance composite variables were created for each rating source by summing ratings on performance dimensions 9–10. In all, six performance criterion composite variables were created and used in sex-based differential prediction analyses: task and contextual performance variables for supervisor, peer, and subordinate ratings. Intraclass correlations were .46, .48, .42, and .44 for peer-rated task performance, peer-rated contextual performance, subordinate-rated task performance, and subordinate-rated contextual performance, respectively. Most participants were rated by only one supervisor, so intraclass correlations could not be calculated for supervisor ratings. Intercorrelations between supervisor, peer, and subordinate ratings of task performance ranged from .18 to .28; and intercorrelations between supervisor, peer, and subordinate ratings of contextual performance ranged from .28 to .38. These low-to-moderate intercorrelations suggest that the ratings from the three sources were independent enough of each other to be considered separately in the present study.

Procedure

Procedures were the same for Samples 1 and 2. Sex-based differential prediction analyses were carried out separately for each rating source (supervisors, peers, and subordinates) and were carried out separately for each pairing of an FFM trait with a performance dimension (i.e. task and contextual performance). This resulted in 30 sex-based differential prediction analyses in both Samples 1 and 2 (5 traits × 3 rating sources × 2 performance dimensions = 30 analyses), for 60 total sex-based differential prediction analyses across the two samples. Step-down hierarchical regression was used for sex-based differential prediction analyses (Lautenschlager & Mendoza, 1986). These analyses test the null hypothesis of equal within-sex regression slopes and intercepts by comparing up to four different nested regression models in stepwise fashion. The four regression models are:

$$\text{Model 1: } \hat{Y} = b_0 + b_1 X + e$$
$$\text{Model 2: } \hat{Y} = b_0 + b_1 X + b_2 S + b_3 XS + e$$
$$\text{Model 3: } \hat{Y} = b_0 + b_1 X + b_3 XS + e$$
$$\text{Model 4: } \hat{Y} = b_0 + b_1 X + b_2 S + e$$

Where $\hat{Y}$ is the predicted criterion score, X is the personality test score, S is a dummy-coded subgroup (i.e. sex) membership variable, XS is a cross-product term obtained by multiplying X and S, and e is a residual. The first step in the step-down analyses is to compare the change in R^2 between Models 1 and 2 as an omnibus test of slope and intercept differences. If the increment in R^2 is not significant, the null hypothesis of equal slopes and intercepts cannot be rejected and no further analyses are appropriate. If the increment in R^2 is significant, this suggests that sex-based differential prediction exists in the form of differences between sexes' regression slopes and/or intercepts. Only in this case are sequential tests of equivalence of slopes and/or intercepts carried out in order to determine the exact form of sex-based differential prediction. First, to test for slope differences, Models 2 and 4 (which only differ in inclusion of the interaction term) are compared. If the change in R^2 is significant, this suggests that there are slope differences between sexes. If there are slope differences between sexes, this makes the intercept differences uninterpretable, and therefore an individual test of intercept differences is not carried out. However, if there is no significant change in R^2 between Models 2 and 4, this suggests that slopes do not differ, and therefore a test of whether intercepts differ between subgroups is appropriate. To test for intercept differences between subgroups, Models 2 and 3 (which differ only in inclusion of the subgroup dummy variable) are compared; a significant change in R^2 suggests a difference between subgroups in intercepts.

RESULTS

Correlations between all study variables are listed in Table 3 for both Samples 1 and 2. Each of the FFM traits exhibited significant correlations with at least some performance ratings. Further, the magnitude of personality–performance relationships in both Samples 1 and 2 are similar to the typical range of observed validities for FFM traits documented in meta-analyses (Barrick et al., 2001).

Means, standard deviations, and standardised mean differences (*d*s) for males and females on each of the FFM traits and job performance ratings are listed in Table 4[2] for both Samples 1 and 2. In Sample 1, although most mean differences were small, women had lower mean scores on each of the FFM

[2] We note that the *T*-scores on each of the personality dimensions for Sample 1 were relatively high (i.e. .50–.80 standard deviations above the average score of 50). The reason for this is not

TABLE 4
Means, Standard Deviations, and Standardised Mean Differences for Each Sex

	Females		*Males*		
	Mean	SD	*Mean*	SD	d
Sample 1					
Extraversion	53.95	5.84	55.00	5.83	−.18
Agreeableness	57.09	7.07	57.55	7.24	−.06
Openness	55.02	5.34	56.46	5.36	−.27*
Conscientiousness	57.46	6.13	57.72	6.59	−.04
Emotional Stability	55.64	5.72	57.95	6.52	−.37*
Task Performance—Supervisor Rating	36.99	5.54	36.18	6.41	.13
Task Performance—Peer Rating	36.66	4.14	36.66	4.22	.00
Task Performance—Subordinate Rating	41.37	4.61	40.83	5.13	.11
Contextual Performance—Supervisor Rating	31.51	5.42	30.65	5.55	.16
Contextual Performance—Peer Rating	30.71	3.57	30.04	3.65	.18
Contextual Performance—Subordinate Rating	33.71	4.68	33.12	5.13	.12
Sample 2					
Extraversion	52.99	6.91	51.97	7.26	.14
Agreeableness	46.15	7.89	44.47	8.37	.20*
Openness	46.95	7.82	49.90	7.36	−.40*
Conscientiousness	48.54	8.30	47.08	8.37	.18
Emotional Stability	49.04	8.96	48.36	8.94	.08
Task Performance—Supervisor Rating	24.88	2.78	24.78	2.89	.03
Task Performance—Peer Rating	24.59	2.12	24.56	2.09	.02
Task Performance—Subordinate Rating	26.05	1.91	26.05	1.86	.00
Contextual Performance—Supervisor Rating	6.17	.78	6.14	.78	.04
Contextual Performance—Peer Rating	6.17	.61	6.19	.58	−.04
Contextual Performance—Subordinate Rating	6.36	.53	6.44	.52	−.15

Note: The task and contextual performance rating instruments were different for Samples 1 and 2 (see Methods section); *d* = standardised mean difference; positive values mean females scored higher; * $p < .05$.

traits, but higher mean scores on most of the job performance ratings, implying the potential for under-prediction of female performance. In Sample 2 there were virtually no sex differences in mean performance ratings, but there

particularly clear. One possibility is socially desirable responding (SDR). However, all else equal, SDR would be expected to attenuate personality validity (e.g. Berry & Sackett, 2009), and Sample 1's personality validities are actually higher than are often seen for personality traits. Another possible explanation is that the participants in Sample 1 were relatively high-level managers. To the degree that personality traits predict job performance, and to the degree that higher performers are more likely to become managers, one would expect managers to score higher than average on positive personality traits. A third possibility arises from research demonstrating that individuals develop in response to their experiences and the expectations of their social roles (Roberts, Caspi, & Moffitt, 2003). It is possible that the experiences and social expectations of high-level managers shape them to be more emotionally stable, extraverted, open to experience, conscientious, and agreeable.

TABLE 5
Unstandardised Regression Coefficients Resulting from the Step-down Hierarchical Regression Analyses for Samples 1 and 2

	EX	*AG*	*OP*	*CO*	*ES*
Sample 1 Sex-Based Differential Prediction Analyses					
Supervisor ratings					
Task Performance	−.17	−.13	−.15	−.13	**.15***
Contextual Performance	−.18	−.15	−.15	−.16	−.20
Peer ratings					
Task Performance	−.07	−.02	−.10	−.03	−.13
Contextual Performance	**.10***	−.19	−.17	−.18	−.26
Subordinate ratings					
Task Performance	−.13	−.13	−.14	−.12	−.17
Contextual Performance	−.11	−.13	−.07	−.12	−.15
Sample 2 Sex-Based Differential Prediction Analyses					
Supervisor ratings					
Task Performance	−.01	.00	−.07	.00	−.01
Contextual Performance	.01	.01	−.05	.01	−.01
Peer ratings					
Task Performance	.00	.01	−.03	.01	−.01
Contextual Performance	.10	.11	.08	.07	.10
Subordinate ratings					
Task Performance	.02	.03	.01	.02	.02
Contextual Performance	.07	.07	.12	.05	.07

Note: When a coefficient is bolded it represents the regression coefficient for the sex–trait interaction term (i.e. slope difference) when performance criteria are regressed on the FFM trait score, the sex dummy variable, and a sex–trait interaction term. When a coefficient is not bolded, it is the regression coefficient for the sex dummy variable (i.e. intercept difference) when performance criteria are regressed on the FFM trait score, the sex dummy variable, and a sex–trait interaction term. An asterisk means the regression coefficient is statistically significant ($p < .05$). EX = Extraversion, AG = Agreeableness, OP = Openness to Experience, CO = Conscientiousness, ES = Emotional Stability.

were some small-to-moderate mean differences on FFM traits (e.g. women scored lower on Openness, men scored slightly lower on Agreeableness), suggesting the possibility for sex-based differential prediction.

Sample 1 Sex-Based Differential Prediction Results

Results of the 30 sex-based differential prediction analyses carried out for Sample 1 are listed in the top half of Table 5. Each number in Table 5 is an unstandardised regression coefficient. Because in the step-down hierarchical regression method a significant slope difference precludes the need to test for intercept differences, intercept difference results are only listed when the slopes did not differ. Therefore, bolded entries represent the regression coef-

ficient for the sex–trait interaction term (i.e. slope difference) when performance criteria are regressed on the FFM trait score, the sex dummy variable, and a sex–trait interaction term. When a coefficient is not bolded, it is the regression coefficient for the sex dummy variable (i.e. intercept difference) when performance criteria are regressed on the FFM trait score, the sex dummy variable, and a sex–trait interaction term (when the sex coefficient is listed, it can be assumed that there were not significant slope differences, and thus only the sex/intercept coefficient is listed). Because the sex dummy variables were coded such that 0 = female and 1 = male, positive intercept coefficients (unbolded coefficients) mean that the male intercept is higher, while negative coefficients mean that the female intercept is higher. Positive slope coefficients (bolded coefficients) mean that the FFM trait is more strongly related to performance for men, while negative slope coefficients mean the FFM trait is more strongly related to performance for women. To aid in interpretation, the performance ratings were standardised such that they had a mean of zero and standard deviation of one, while the predictor variables (sex and FFM traits) were not standardised. Thus, positive unstandardised regression coefficients for intercept differences in Table 5 can be directly interpreted as the number of criterion standard deviation units higher the male intercept was than the female intercept (with the opposite being true of negative coefficients).

Across the 30 analyses, there were two instances of sex-based differential prediction, meaning that the slopes or intercepts differed for sexes 6.7 per cent of the time. This is approximately equivalent to the alpha level of .05, suggesting that sex-based differential prediction was essentially a chance phenomenon in Sample 1. It is interesting to note that of the two instances of sex-based differential prediction, both were slope differences between sexes. Both slope coefficients were positive, meaning that Emotional Stability was more strongly related to supervisor ratings of task performance for men and Extraversion was more strongly related to peer ratings of contextual performance for men. Emotional Stability and Extraversion are relatively agentic traits, which is in line with Hypothesis 1 that suggested that any instances of sex-based differential prediction would be more likely to occur with relatively agentic FFM traits. However, we hesitate to call this support for Hypothesis 1, as slope differences occurred hardly more often than would be expected due to chance. There was no support for Hypotheses 2 or 3 that suggested that sex-based differential prediction would be more likely for task performance and supervisor ratings, respectively.

Sample 2 Sex-Based Differential Prediction Results

Results for the Sample 2 sex-based differential prediction analyses are listed in the bottom half of Table 5. Across the 30 analyses, there were no instances

of sex-based differential prediction, meaning that the slopes or intercepts did not differ across sexes. There was no support for Hypotheses 1, 2, or 3. Thus, similar to the results for Sample 1, sex-based differential prediction does not appear to be an issue in Sample 2.

DISCUSSION

Summary of Findings

Across two separate samples, a comparably low prevalence of sex-based differential prediction was identified (6.7% and 0% of the time in Samples 1 and 2, respectively). Out the 60 sex-based differential prediction analyses across Samples 1 and 2, there were only two instances of sex-based differential prediction, meaning that sex-based differential prediction only occurred 3.3 per cent of the time. This is even less than would be expected due to chance alone, given an alpha level of .05 (i.e. across 60 analyses, the alpha level of .05 suggests that as many as three significant results could be found due to chance). This suggests that sex-based differential prediction is a phenomenon that occurs no more often than would be expected by chance alone in the personality–performance domain.

Therefore, no evidence was found that FFM traits are more strongly related to supervisor, peer, or subordinate ratings of job performance for men than for women, suggesting that inferences drawn from personality test scores are equally valid for each sex. Also, no evidence was found that FFM traits under-predict (or over-predict) performance ratings for women, suggesting that the predictive meaning of FFM trait scores is equivalent for men and women. Combined with the results of Saad and Sackett's (2002) military study, a body of evidence for a lack of sex-based differential prediction of performance for personality measures is beginning to accumulate; although this is only based on two studies. This is an important applied finding as it bolsters the position of personality tests as a relatively fair and unbiased tool for use in organisations.

Comparisons with Results from Previous Research

The current study was an initial step toward addressing the lack of research on sex-based differential prediction in employment settings. The only study to date that has investigated sex-based differential prediction was Saad and Sackett's (2002) military sample study. Overall, the general message from the current study was very similar to that from Saad and Sackett, with results only differing in some specific details. First, the big picture message from both the current study and Saad and Sackett was the same: both studies found no evidence compatible with personality test bias. Second, neither the

current study nor Saad and Sackett identified slope differences between sexes at greater-than-chance levels. The main difference in the pattern of results between the two studies is that Saad and Sackett found evidence of intercept differences while the present study found none. Thus, the picture that begins to emerge is one of some sex-based differential prediction in the form of intercept differences, but a lack of evidence of slope differences or personality test bias. This is remarkably similar to the pattern of results typically observed for cognitive ability tests (Schmidt, 1988).

The difference between Saad and Sackett (2002) and the current study in intercept differences should be discussed. Saad and Sackett found that women had a lower intercept in about one-third of analyses while the current study found no evidence of intercept differences. Because intercept differences in sex-based differential prediction represent differences between sexes in performance unaccounted for by either differences (a) between sexes in test scores or (b) in the relationship between test scores and performance, intercept differences are most likely to occur when there are criterion score differences between sexes (e.g. if women score lower on performance, they are more likely to have the lower intercept, meaning that female performance would be over-predicted by the common regression line). Not surprisingly, in Saad and Sackett's military sample, men scored higher than women on most performance dimensions. In the present study, where there were criterion score differences, they were most often in favor of women. Therefore, the explanation for the differences regarding intercept findings between the present study and Saad and Sackett's may reside in the mechanism causing the opposite direction of criterion score differences.

There were at least three relevant differences in the design of Saad and Sackett (2002) versus the present study: the personality tests (the ABLE vs. the WBI, respectively), the criteria (supervisor ratings, work samples, and personnel files vs. supervisor, peer, and subordinate ratings, respectively), and the setting (military vs. civilian, respectively). Both personality tests (the ABLE and WBI) were designed to measure similar personality traits, and it is not clear to us why properties of the tests would influence criterion score differences between the studies; so, we think differences between studies in personality tests are an unlikely explanation for the differences in intercept findings. Although most of the performance criteria differed between the two studies, both Saad and Sackett and the present study included supervisor ratings and the intercept difference findings still differed between the studies on this common criterion. This argues against differences in the criteria across the two studies explaining the differences in intercept findings. We believe a more likely explanation for the disparity between studies arises from the use of military versus civilian managerial samples. In a setting as historically dominated by men as the Army, it is not surprising that women would, on average, have lower performance scores, resulting in conditions ripe for

over-prediction of performance. Although managerial settings have historically been similarly dominated by men (e.g. Gutek, 1993; Rinfret & Lortie-Lussier, 1996), recent research suggests that this is less the case today (Duehr & Bono, 2006) and that women tend to exhibit some positive managerial behaviors, such as transformational leadership, slightly more often than men (Carless, 1998). Thus, sizable mean performance differences between sexes are less likely in managerial samples, making it less likely that intercept differences will occur. This was indeed the case in the present study as there were not sizable mean differences in performance between men and women. This highlights the importance of attending to the job type when investigating sex-based differential prediction.

Practical Implications

The use of personality testing for purposes such as personnel selection/screening or employee development is common in organisations (Daniel, 2005). Thus, it is important that personality tests not be biased against important subgroups, especially those subgroups outlined in Title VII of the Civil Rights Act. Women are one such subgroup, and the results of the current study, along with the results of Saad and Sackett (2002), suggest that personality tests do not exhibit predictive bias against women. This finding must be replicated in other jobs, organisations, and for other personality tests. However, to date no sex-based differential prediction studies have identified evidence of personality test bias. This suggests that, unless findings change in future research using personality tests other than the WBI and ABLE (as were used in the present study and Saad and Sackett, 2002), organisations using personality tests do not have reason for concern.

Additional Issues, Limitations, and Directions for Future Research

The present study was not without limitations. One potential limitation is the size of the female samples (60 and 134 in Samples 1 and 2, respectively), which affects statistical power. In moderated multiple regression analyses, of which differential prediction analyses are one type, statistical power to detect interactions (i.e. slope differences) is probably the greatest issue (Aguinis, Beaty, Boik, & Pierce, 2005). A computer program called MMRPower described by Aguinis, Boik, and Pierce (2001) was used to determine the statistical power that the present study had to detect slope differences of moderate size (Cohen, 1992). Depending on the specific predictor–criterion combination, statistical power in Sample 1 ranged from .46 to .57, while statistical power in Sample 2 ranged from .85 to .93. Thus, statistical power to detect interactions was not ideal in Sample 1. However, the only slope

differences detected in either sample were in Sample 1, the sample with less power. Therefore, it is unlikely that there were actually sizable slope differences in the present study that went undetected due to low power. Regardless, future research with larger samples is warranted.

The use of managerial samples represented both a strength and a limitation of the current study. On the one hand, the only previous sex-based differential prediction study (Saad & Sackett, 2002) used a military sample, and thus the use of any employment sample in the current study acted as a needed extension of previous research. Also, unlike most other employees, managers have supervisors, peers, and subordinates, which allowed us to investigate whether patterns of results differed across these three rating sources. On the other hand, the use of managerial samples in the current study does not allow us to generalise with confidence to other civilian jobs. Future sex-based differential prediction research should be carried out using employees in other types of jobs.

The current study used only one personality instrument: the Work Behavior Inventory (WBI). Thus, the degree to which the results of the present study generalise to other personality inventories is still an open question in need of future research. There are, however, two points that speak to this issue. For one, past research has demonstrated that scores on WBI scales correlate highly with scores on relevant scales of other commonly used personality instruments such as the Hogan Personality Inventory, the Occupational Personality Questionnaire, the Golden Personality Type Indicator, and the BarOn EQ-i (Page, 2009). Thus, it is unlikely there is something idiosyncratic about the WBI. Second, the pattern of results in the current study was very similar to the pattern of results for the personality instrument used in Saad and Sackett (2002; US Army's Assessment of Background and Life Experiences). This acts as further convergent evidence. Regardless, future sex-based differential prediction research using different FFM personality measures, or even measures of traits outside of the FFM (e.g. Lee, Ashton, & DeVries, 2005), would be worthwhile.

The present study did not have access to information about the individual performance raters. Thus, it was not possible to investigate the degree to which characteristics of individual raters (e.g. sex, age, endorsement of sex stereotypes) might influence the results. It is possible that personality trait scores might have differentially predicted performance for certain types of performance raters, such as male raters, or raters that endorse sex stereotypes. We suggest this as a direction for future research.

An additional issue regards the two different cultures from which Sample 1 and Sample 2 were drawn. Sample 1 was an American sample while Sample 2 was a Chinese sample. As previously mentioned, a number of studies have documented the comparability of the WBI for Chinese and American samples (Liang & Yang, 2006; Page, 2009; Thompson et al., 2008), suggest-

ing that the meaning of the personality measure should not have differed across samples. Less clear is whether the meaning of the criterion measures was comparable across the two samples. It is noteworthy that similar factor structures of performance ratings were found in each sample. However, this does not necessarily mean that the performance raters interpreted the meaning of performance dimensions for men and women similarly across the samples. Regardless, the general pattern of results (i.e. no intercept differences, almost no slope differences, no consistent evidence of personality test bias) was the same across the two samples, making it less likely that cultural differences confound major study conclusions. In fact, that the general pattern of results replicated across two different cultures might be considered a strength of the current study.

REFERENCES

Aguinis, H., Beaty, J.C., Boik, R.J., & Pierce, C.A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *90*, 94–107.

Aguinis, H., Boik, R.J., & Pierce, C.A. (2001). A generalized solution for approximating the power to detect effects of categorical moderator variables using multiple regression. *Organizational Research Methods*, *4*, 291–323.

Ash, P., Baehr, M.E., Joy, D.S., & Orban, J.A. (1988). Employment testing for the selection and evaluation of bus drivers. *Applied Psychology: An International Review*, *37*, 351–363.

Barrick, M.R., & Mount, M.K. (1991). The Big Five Personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1–26.

Barrick, M.R., Mount, M.K., & Judge, T.A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment*, *9*, 9–30.

Berry, C.M., Ones, D.S., & Sackett, P.R. (2007a). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology*, *92*, 410–424.

Berry, C.M., Page, R.C., & Sackett, P.R. (2007b). Effects of self-deceptive enhancement on personality–job performance relationships. *International Journal of Selection and Assessment*, *15*, 94–109.

Berry, C.M., & Sackett, P.R. (2009). Faking in personnel selection: Tradeoffs in performance versus fairness resulting from two cut-score strategies. *Personnel Psychology*, *62*, 835–863.

Bobko, P., Roth, P.L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology*, *52*, 561–589.

Borman, W., & Motowidlo, S. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance*, *10*, 99–109.

Burgess, D., & Borgida, E. (1999). Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, Public Policy, and Law*, *5*, 665–692.

Carless, S.A. (1998). Sex differences in transformational leadership: An examination of superior, leader, and subordinate perspectives. *Sex Roles*, *39*, 887–902.

Chiaburu, D.S., Oh, I.-S., Berry, C.M., Li, N., & Gardner, R.G. (2011, June 20). The five-factor model of personality traits and organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology*, *96*, 1140–1166.

Cleary, T.A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, *5*, 115–124.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Conway, J.M., & Huffcutt, A.I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, *10*, 331–360.

Costa, P.T. (1996). Work and personality: The use of the NEO-PI-R in industrial/organizational psychology. *Applied Psychology: An International Review*, *45*, 225–241.

Daniel, L. (2005, April–June). Use personality tests legally and effectively. *Staffing Management*, *1*. Retrieved 27 February 2012 from http://www.shrm.org/Publications/StaffingManagementMagazine/EditorialContent/Pages/0504_cover.aspx.

Davison, H.K., & Burke, M.J. (2000). Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior*, *56*, 225–248.

Duehr, E.E., & Bono, J.E. (2006). Men, women, and managers: Are stereotypes finally changing? *Personnel Psychology*, *59*, 815–846.

Eagly, A.H., Makhijani, M.G., & Klonsky, B.G. (1992). Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin*, *111*, 3–22.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*, 429–456.

Gutek, B.A. (1993). Changing the status of women in management. *Applied Psychology: An International Review*, *42*, 301–311.

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory Manual*. Tulsa, OK: Hogan Assessment Systems.

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job–performance relations: A socioanalytic perspective. *Journal of Applied Psychology*, *88*, 100–112.

Hough, L.M., Oswald, F.L., & Ployhart, R.E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment*, *9*, 152–194.

Hurtz, G.M., & Donovan, J.J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology*, *85*, 869–879.

Johnson, J. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. *Journal of Applied Psychology*, *86*, 984–996.

Judge, T.A., Bono, J.E., Ilies, R., & Gerhardt, M.W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology*, *87*, 765–780.

Lautenschlager, G.J., & Mendoza, J.L. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement*, *10*, 133–139.

Lee, K., Ashton, M.C., & de Vries, R.E. (2005). Predicting workplace delinquency and integrity with the HEXACO and five-factor models of personality structure. *Human Performance*, *18*, 179–197.

Liang, K.G., & Yang, X. (2006, May). Personality assessment in Chinese firms. In R.C. Page (Chair), *Cross cultural barriers: Validity of personality assessment in non-Western cultures*. Symposium at the 2006 Annual Conference of the Society for Industrial and Organizational Psychology, Dallas TX.

London, M., & Smither, J.W. (1995). Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance related outcomes? Theory-based applications and directions for research. *Personnel Psychology*, *48*, 803–839.

Lyness, K.S., & Heilman, M.E. (2006). When fit is fundamental: Performance evaluations and promotions of upper-level female and male managers. *Journal of Applied Psychology*, *91*, 777–785.

Motowidlo, S., Borman, W., & Schmit, M. (1997). A theory of individual differences in task and contextual performance. *Human Performance*, *10*, 71–83.

Motowidlo, S., & Van Scotter, J. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology*, *79*, 475–480.

Oh, I., & Berry, C.M. (2009). The five-factor model of personality and managerial performance: Validity gains through the use of 360 degree performance ratings. *Journal of Applied Psychology*, *94*, 1498–1513.

Ones, D.S., & Viswesvaran, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior*, *17*, 609–626.

Ones, D.S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance*, *11*, 245–269.

Page, R.C. (2009). *Page Work Behavior Inventory: Manual & user's guide* (2nd edn.). Minneapolis, MN: Assessment Associates International.

Paul, A.M. (2004). *The cult of personality: How personality tests are leading us to miseducate our children, mismanage our companies, and misunderstand ourselves*. New York: Free Press.

Paulhus, D.L., & John, O.P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, *66*, 1025–1060.

Peterson, N.G., Hough, L.M., Dunnette, M.D., Rosse, R.L., Houston, J.S., Toquam, J.L., & Wing, H. (1990). Project A: Specification of the predictor domain and development of new selection/classification tests. *Personnel Psychology*, *43*, 247–276.

Rinfret, N., & Lortie-Lussier, M. (1996). Changes in attitudes towards female managers: Model differences as a function of sex. *Applied Psychology: An International Review*, *45*, 353–370.

Roberts, B.W., Caspi, A., & Moffitt, T. (2003). Work experiences and personality development in young adulthood. *Journal of Personality and Social Psychology*, *84*, 582–593.

Rudman, L.A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology*, *74*, 629–645.

Rudman, L.A., & Fairchild, K. (2004). Reactions to counterstereotypic behavior: The role of backlash in cultural stereotype maintenance. *Journal of Personality and Social Psychology*, *87*, 157–176.

Saad, S., & Sackett, P.R. (2002). Investigating differential prediction by sex in employment-oriented personality measures. *Journal of Applied Psychology*, *87*, 667–674.

Sackett, P.R., Borneman, M.J., & Connelly, B.S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, *63*, 215–227.

Sackett, P.R., & Wilk, S.L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, *49*, 929–954.

Saville, P., Sik, G., Nyfield, G., Hackston, J., & MacIver, R. (1996). A demonstration of the validity of the Occupational Personality Questionnaire (OPQ) in the measurement of job competencies across time and in separate organizations. *Applied Psychology: An International Review*, *45*, 243–262.

Schmidt, F.L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior*, *33*, 272–292.

Sheppard, R., Han, K., Colarelli, S.M., Dai, G., & King, D.W. (2006). Differential item functioning by sex and race in the Hogan Personality Inventory. *Assessment*, *13*, 442–453.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th edn.). Bowling Green, OH: Author.

Thompson, A., Hartmann, L., Vang, M., & Tubré, T. (2008). Alternate-forms reliability assessment of a Five-Factor-Model personality inventory across Mandarin Chinese and English bilingual speakers. Poster presented at the 2008 Minnesota Psychological Association Annual Convention, Bloomington, MN.

Van Scotter, J.R., & Motowidlo, S.J. (1996). Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of Applied Psychology*, *81*, 525–531.

Wang, Y., Fang, M., & Mobley, W.H. (2006). *Manual and interpretation guide for MGP 360-degree leadership survey*. Shanghai, China: Mobley Group Pacific.

APPENDIX

TABLE A1
Example Items for each of the Five-Factor Model Traits in the Work Behavior Inventory

	Example items
Extraversion	"I find it easy to meet people and make new friends" "People would describe me as shy"—R
Agreeableness	"I cooperate well with other team members" "I am not very patient at listening to the concerns of others"—R
Openness	"I am good at coping with change" "I usually don't have many creative ideas"—R
Conscientiousness	"I make a habit of double checking my work for accuracy" "I don't like working toward difficult goals"—R
Emotional Stability	"I rarely lose my temper" "I worry a lot"—R

R = Reverse-coded