Journal of Research in Personality 46 (2012) 8–18

Contents lists available at SciVerse ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Not too little, but not too much: The perceived desirability of responses to personality items ☆

Patrick D. Dunlop *, Amelia D. Telford, David L. Morrison

School of Psychology, University of Western Australia, Australia

ARTICLE INFO

Article history:
Available online 21 October 2011

Keywords:
Personality
Response desirability
Impression management
Faking
High stakes
Likert scale
HEXACO

ABSTRACT

Paradigms typically employed to investigate socially desirable responding in personality assessment implicitly assume linear relationships exist between trait level and desirability but recent research has called this assumption into question. In this study, participants rated the desirability of a hypothetical applicant to one of four jobs on the basis of which five-point Likert-type scale option he/she selected when responding to personality items. Results generally indicated that the most extreme option, on the desirable side of the response scale, was rated as most desirable, but perceived desirability asymptotes with the penultimate option. The middle (neutral) option, however, was consistently regarded as being much less desirable. The occupational context also significantly moderated the patterns of desirability ratings for many items.

1. Introduction

One commonly cited concern around the use of self-report personality assessments in high-stakes settings, such as personnel selection, is their alleged susceptibility to socially desirable responding, or 'faking' (e.g. Morgeson et al., 2007; Rothstein & Goffin, 2006). Indeed, it is a well-established finding in the personnel selection literature that job applicants tend to score higher than non-applicants on scales measuring personality traits that are typically most predictive of organizationally-relevant criteria (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006). Socially desirable responding in high-stakes settings has therefore often thought to emerge as a process whereby individuals first identify certain traits as being desirable in relation to the context at hand and, consequently, endorse items measuring these traits in their self-assessments. Some recent research (e.g. Borkenau, Zaltauskas, & Leising, 2009; Kuncel & Tellegen, 2009) suggests, however, that *extreme* levels of ostensibly 'desirable' personality traits may not be perceived as desirable, thus calling into question the assumption that socially desirable responding will be synonymous with extreme endorsement.

Researchers have long been aware of the positive association between the perceived social desirability of a self-descriptive statement and the probability of it being endorsed by an individual (e.g. Edwards, 1953). Endorsement of an item by an individual is relatively straightforward when faced with a dichotomous response scale (e.g. True/False, Yes/No). Ordered polytomous response scales (e.g. Strongly Disagree, Disagree, Neither, Agree, Strongly Agree), which are used in many personality assessments, add an additional layer of complexity, however, in that individuals must also make decisions regarding the *extent* to which they endorse the item. Nonetheless, traditional *statistical* conceptualizations of the social desirability of personality items (e.g. those which treat desirability as an inherent property of an item; e.g. Edwards, 1953; Pauls & Crost, 2005), have implicitly assumed that there is a linear relationship between desirability and the response options. Thus, Strongly Agreeing with an item measuring a desirable trait is considered a more socially desirable response than merely Agreeing. By extension, an individual who wants to falsely portray a maximally socially desirable profile would presumably Strongly Agree with all items measuring desirable traits, and Strongly Disagree with all items measuring undesirable traits (Snell, Sydell, & Lueke, 1999).

Two recent papers by Borkenau et al. (2009) and Kuncel and Tellegen (2009) have suggested, however, that such a presumption may be unwarranted. In their social relations study, Borkenau et al. found evidence of non-linear relationships between trait levels and perceived social desirability. Borkenau et al. asked participants to directly indicate the level, on a set of six-point bipolar trait scales, which they felt was most desirable. It was generally the case for

☆ An earlier version of this paper was presented at the 26th Annual Society for Industrial and Organizational Psychologists Conference, 2011, Chicago, IL, USA.

* Corresponding author. Address: School of Psychology, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia. Fax: +61 8 6488 1006.

E-mail address: patrick.dunlop@uwa.edu.au (P.D. Dunlop).

0092-6566/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jrp.2011.10.004

the bipolar scales that one pole was considered distinctly more desirable than the other. Overwhelmingly, however, the *level* of the trait considered most desirable was that captured by the penultimate, and not the extreme, point on the response scale.

Whilst Borkenau et al.'s (2009) study provides evidence of curvilinear relationships between trait levels and desirability, one concern might be that such a pattern is perhaps peculiar to bipolar response scales (see Yorke, 2001 for a critical review of bipolar response scales). Nonetheless, using unipolar adjectival ratings, Kuncel and Tellegen (2009) also found evidence of non-linear relationships between trait level and perceived desirability. In Kuncel and Tellegen's first study, participants were asked to rate how desirable a person would be if he or she was extremely high (top 1%), above average (top 30%), average, below average (bottom 30%), or extremely low (bottom 1%) on each of a set of self-descriptive adjectives (e.g. talkative, conservative). Using these ratings, Kuncel and Tellegen plotted 'desirability functions' (i.e. mean desirability ratings plotted against the five trait levels) for each item. Whilst a large number of the adjectives did exhibit desirability functions that were roughly linear, these adjectives mainly captured the highly evaluative traits Positive and Negative Valence. Such traits are arguably expected to relate linearly to desirability (see also Borkenau et al., 2009). As it happens, however, items capturing these traits are rarely incorporated in questionnaires used in high-stakes selection contexts and therefore these results, while interesting, may not generalize to such settings. By contrast, almost all of the adjectives which assessed personality traits within the Big Five space yielded desirability functions with a turning point (inflection). For some, the turning point was at the center, producing an inverted U-shaped function; that is, it is considered most desirable to be average on that trait and less desirable to be at either extreme. For most adjectives, however, the turning point was at the above average position, indicating that high levels of the trait were more desirable, but only to a point; a result consistent with Borkenau et al.'s findings.

Borkenau et al. (2009) and Kuncel and Tellegen's (2009) studies both challenge the assumption that social desirability is linearly related to trait levels. Nonetheless, the adjectival rating method is relatively uncommon in personality questionnaires used in high-stakes settings when compared to the more frequently adopted full-statement rating method. Further, polytomous response scales often utilized in practice tend to comprise subjective anchors (e.g. Strongly Disagree, Strongly Agree), whereas the two studies above employed, respectively, bipolar scales and objective points on a normative continuum to define trait levels. The present study therefore employs a similar methodology to that of Kuncel and Tellegen's (2009) first study but it directly addresses the limitations described above by asking participants to consider points on a subjective Strongly Disagree–Strongly Agree Likert-type scale against full statement personality items. As a second extension of previous studies, participants in this study were also asked to consider the desirability of the different statements, against the context of an individual applying for either a 'general' job or for one of three specific jobs: fire fighter, nurse, or car salesperson. The aim here was to assess the impact of context on the relationship between trait levels and perceived desirability. Potential implications for these methodological innovations are articulated below.

It is not immediately clear what the impact of asking participants to rate the desirability of the Strongly Disagree–Strongly Agree response options will be. On the one hand, this response scale requires individuals to make their own subjective judgments about how the response anchors correspond to different levels of the traits being measured. So whilst it may be true that being more talkative, for example, than 99% of the population is seen as undesirable, it does not necessarily follow that selecting Strongly Agree in response to a statement measuring talkativeness is akin to claiming to be more talkative than 99% of the population. Expressed another way, the extreme subjective response options (i.e. Strongly Agree and Strongly Disagree) may not necessarily be seen to reflect truly extreme levels of the underlying traits. To the extent that this is true, the impact is likely to be that the declines in perceived desirability at the extreme trait levels observed by Kuncel and Tellegen (2009) will not emerge here. On the other hand, Kuncel and Tellegen's second study, in which participants under directed-faking conditions were asked to explain why they did not select the Strongly Agree option for an item, suggested that many people recognized that this option *was* extreme, often believing it to be undesirable. Furthermore, the response scale used by Borkenau et al. (2009) makes no reference to normative comparisons either, yet they still observed non-linear trait level-desirability relationships. With reasonable arguments on both sides, this element of the study was approached with an exploratory mindset.

Past research has suggested that individuals can adapt their impression management response strategy to suit the context (Furnham, 1990; Krahe, 1989; Mahar, Cologon, & Duck, 1995; Mahar et al., 2006; Pauls & Crost, 2005), indicating that they are sensitive to the varying contextual intricacies. We therefore hypothesize that the occupational context would impact upon the perceived desirability of different levels of some personality traits. To make specific predictions about exactly how the occupational context might impact on the perceived desirability of different trait levels, however, it is important to first consider the content captured by the personality model being explored, which in this case was the HEXACO model (Ashton & Lee, 2001, 2007). One can imagine how the content captured by the HEXACO framework might be of particular relevance to the different occupational contexts under study, namely car salesperson, fire fighter, and nurse. First, being a successful car salesperson will require an incumbent to be comfortable approaching, talking to, and negotiating with strangers, hence higher levels of Extraversion seem likely to be perceived as being especially desirable in this context. Further, high levels of Honesty–Humility might also be regarded as undesirable for this role as being too sincere, modest and avoidant of overt signs of financial success may impinge on one's ability to quickly build credibility with customers. Second, being a successful fire fighter may be seen as requiring an incumbent to face potentially dangerous or traumatic situations whilst maintaining a calm demeanor. For this occupation, low levels of Emotionality as captured within HEXACO by content on fearfulness and anxiety would appear likely to be especially desirable traits for an aspiring fire fighter. Lastly, being a nurse is likely to require an incumbent to be highly sensitive to the needs and ails of others. Consequently, high levels of sentimentality, as captured within the Emotionality factor, and gentleness and patience, as captured within the Agreeableness factor may be perceived as being particularly desirable for aspiring nurses.

Whilst the context is predicted to moderate the perceived desirability of different levels of some personality traits, there are likely to be other personality traits that *are* of equal relevance or desirability to many jobs. For example, given its job-relevance, one would expect that the perceived desirability of different levels of Conscientiousness is likely to be fairly consistent across all of the occupational contexts considered in this study. Further, Openness is often thought of as the factor that gets manipulated the least in high-stakes situations (though see Birkeland et al., 2006; Griffin, Hesketh, & Grayson, 2004), presumably because it is not seen as being either desirable or undesirable. We therefore expected that the different levels of the Openness trait would be considered equally desirable across all contexts.

To summarize, the aim of this study is to examine the generalizability of the Borkenau et al. (2009) and Kuncel and Tellegen

(2009) findings and we hypothesize that social desirability and item response options are not linearly related, as is generally assumed. A second aim was to demonstrate the sophisticated nature of personality questionnaire response patterns by illustrating context sensitivity of socially desirable responding.

2. Method

2.1. Participants

A total of 239 individuals participated in this study. Participants were recruited by the authors via email, word of mouth, and a social networking website. A 'snowball' recruiting method was also employed whereby participants were asked to recruit additional people that they knew. Due to participants being able to complete the questionnaire anonymously, it was not possible to determine how many were recruited in this manner. Although information on nationality was not collected, it is likely that almost all participants were Australian. The mean age was 33.49 years (SD = 14.83, range: 19–68) and 65% were female. Participants were asked if they had ever completed a personality questionnaire in relation to an application to a job, and 28% reported that they had. No incentives were offered for participation.

2.2. Design

This study employed a mixed factorial design with the between-group factor, "Vignette", representing the type of job description and selection criteria that participants were presented with: general job, fire fighter, nurse, or used car salesperson. The personnel selection scenario was chosen in this case as it provided a context where, in practice, the problem of socially desirable responding is thought to be widespread (e.g. Morgeson et al., 2007; Rothstein & Goffin, 2006). The three specific jobs were chosen because they were expected to cover a wide variety of personality dimensions which are characteristic of the 'ideal' employee. The within-subjects factor was labeled "Response Option" with five levels that reflected the five options of a Strongly Disagree–Strongly Agree Likert scale.

2.3. Materials

2.3.1. Vignettes and job descriptions

Three vignettes were developed for the fire fighter, nurse, and used car salesperson Vignette conditions (refer to Appendix A for the content of the vignettes). Each provided a brief description of the corresponding job and assessment criteria and were formulated on the basis of current print and online job advertisements. Although there was also a 'general job' Vignette group, no job description or selection criteria were developed for this condition.

2.3.2. Personality items

The present study used a subset of 60 items from Lee and Ashton's (2004) HEXACO Personality Inventory (HEXACO-PI) as a source of personality items (note that the HEXACO model and its associated inventory has since been revised with content on Social Self-Esteem replacing content on Expressiveness; see Ashton & Lee, 2007). This particular subset of items was selected previously by Lee and Ashton (2005) to form a shorter version of the HEXACO-PI, on the basis of the items' relatively higher item-total correlations. The chosen subset of items comprises six 10-item scales designed to measure the personality dimensions Honesty–Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness. Each scale includes a mix of positively and negatively-keyed self-descriptive statements and the items were selected to ensure that each factor scale was represented by the full breadth of its intended content. Example items include "In social situations, I'm usually the one who makes the first move" and "When someone I know well is unhappy, I can almost feel that person's pain myself."

2.4. Procedure

Participants were randomly assigned a priori to one of the four Vignette groups, though due to attrition, the group sizes were only approximately equal (general job $n = 61$, fire fighter $n = 62$, nurse $n = 58$, and used car salesperson $n = 58$). After receiving login details, participants accessed a website which hosted the materials. Upon logging in, participants were informed that a hypothetical person, "J", had applied for a job, as determined by the Vignette condition. Participants were then presented with the corresponding job description and selection criteria (where applicable) and were asked to rate, for each of the 60 items, how desirable they believed "J" would appear as an applicant on the basis of whether he/she had "Strongly Disagreed", "Disagreed", "Neither Agreed nor Disagreed", "Agreed", or "Strongly Agreed" with the item. Thus, each participant provided five ratings for each personality item (i.e. one rating for each response option for each of 60 items) using a five-point response scale: 1 (Very Undesirable) to 5 (Very Desirable). Where applicable, participants were able to refer back to the job description at any time. Appendix B shows the instructions with which the participants were provided.

3. Results

3.1. Preliminary analyses

Prior to conducting substantive analyses, two series of independent samples *t*-tests were conducted to determine whether gender or past experience with a personality questionnaire affected the desirability ratings. Upon doing so, only 11 desirability ratings out of 300 (averaged across the four Vignette groups) exhibited gender differences in ratings at a significance level of $\alpha = .01$, suggesting that gender differences were generally quite trivial. Further, past experience with a personality questionnaire did not affect any of the 300 desirability ratings at a significance level of $\alpha = .01$.

3.2. Substantive analyses

To commence the substantive analyses, data were subjected to a series of mixed 4 (Vignette) by 5 (Response Option) ANOVAs, one for each of the 60 items. A small number of missing responses were observed for several participants in each condition, meaning that the degrees of freedom varied slightly across items. This approach was chosen in favor of a multivariate ANOVA followed by post hoc tests with Bonferroni corrections as the within-item comparisons of desirability ratings were all planned in advance. Nonetheless, due to the sheer volume of statistical tests, it was deemed prudent to apply a conservative critical value of $\alpha = .01$ so as to avoid over-capitalization on chance. The Greenhouse–Geisser correction was also applied to all omnibus tests as nearly all analyses failed Mauchly's test of sphericity of variance, with epsilon values frequently falling below .50.

The effect sizes for the main effects and interaction terms, as quantified by η_p^2, are presented in Table 1. Due to copyright restrictions, item content is replicated for only a subset of 38 items that are within the public domain. For the remaining items, the key theme represented by the item is shown in parentheses instead. A significant main effect of Response Option was observed for all

Table 1
Effect sizes from the mixed ANOVAs and desirability function plots for all items of the 60-item version of the HEXACO-PI.

Factor scale and underlying content	Item (or key theme)	Main effect η_p^2	Interaction η_p^2	Desirability function category by vignette			
				F	N	CS	G
Honesty–Humility							
Sincerity	I wouldn't use flattery to get a raise or promotion at work, even if I thought it would succeed	.192*	.040*	B	B	A	B
Sincerity (R)	(Courts favors by calling on past debts)	.684*	.117*	B	B	B	B
Fairness (R)	If I knew that I could never get caught, I would be willing to steal a million dollars	.695*	.022	B	B	B	B
Fairness (R)	I would be tempted to buy stolen property if I were financially tight	.682*	.011	B	B	B	B
Fairness	(Will pay taxes)	.513*	.007	B	C	B	B
Greed avoidant	Having a lot of money is not especially important to me	.082*	.101*	B	C	A	A
Greed avoidant (R)	(Takes risks to improve status)	.428*	.242*	B	B	A	B
Modesty (R)	(Deserving of authority and influence)	.495*	.081*	B	B	C	B
Modesty	I am an ordinary person who is no better than others	.058*	.068*	B	C	A	A
Modesty (R)	I want people to know that I am an important person of high status	.485*	.086*	B	B	B	B
Emotionality							
Fearfulness (R)	(Is resilient to physical pain)	.233*	.272*	B	A	C	C
Fearfulness	When it comes to physical danger, I am very fearful	.294*	.178*	B	C	A	A
Fearfulness	Even in an emergency I wouldn't feel like panicking	.585*	.102*	B	B	B	C
Anxiety (R)	I worry a lot less than most people do	.234*	.020	B	C	B	C
Anxiety (R)	(Remains calm)	.819*	.087*	B	B	B	B
Dependence (R)	(Requires emotional support)	.385*	.020	B	B	B	C
Dependence	When I suffer from a painful experience, I need someone to make me feel comfortable	.056*	.033	C	A	A	A
Sentimentality	I feel like crying when I see other people crying	.144*	.022	B	C	A	A
Sentimentality	When someone I know well is unhappy, I can almost feel that person's pain myself	.193*	.018	C	C	C	C
Sentimentality (R)	I remain unemotional even in situations where most people get very sentimental	.015	.046*	C	A	A	C
Extraversion							
Expressiveness (R)	(Has a low-key style)	.073*	.035*	A	A	C	A
Expressiveness	(Does more talking than others)	.023*	.228*	A	C	B	C
Social boldness	In social situations, I'm usually the one who makes the first move	.643*	.113*	B	B	B	B
Social boldness	When I'm in a group of people, I'm often the one who speaks on behalf of the group	.494*	.156*	C	C	B	C
Social boldness (R)	(Reluctant to speak out)	.459*	.021	B	B	B	B
Sociability	I prefer jobs that involve active social interaction to those that involve working alone	.622*	.118*	B	B	B	C
Sociability	The first thing that I always do in a new place is to make friends	.751*	.049*	B	B	B	B
Liveliness	(Is a spirited person)	.643*	.077*	B	C	B	B
Liveliness (R)	Most people are more upbeat and dynamic than I generally am	.605*	.040*	B	B	B	B
Liveliness (R)	(Lacks enthusiasm)	.679*	.022	B	B	B	B
Agreeableness							
Forgivingness	I rarely hold a grudge, even against people who have badly wronged me	.643*	.008	B	B	B	B
Forgivingness	(Forgives others)	.638*	.028	B	B	B	B
Gentleness	(Withholds overt criticism)	.450*	.016	B	B	B	B
Gentleness (R)	People sometimes tell me that I am too critical of others	.567*	.014	B	B	B	B
Gentleness	I tend to be lenient in judging other people	.355*	.003	B	C	B	C
Flexibility (R)	People sometimes tell me that I'm too stubborn	.396*	.094*	B	B	A	B
Flexibility (R)	I find it hard to compromise with people when I really think I'm right	.496*	.008	B	B	B	B
Patience (R)	(Easily angered)	.728*	.018	B	B	B	B
Patience (R)	People think of me as someone who has a quick temper	.807*	.021	B	B	B	B
Patience	(Rarely gets angry)	.315*	.019	B	B	B	C
Conscientiousness							
Organization	(Organizes belongings)	.714*	.038*	B	B	B	C
Organization	(Returns items when finished)	.733*	.044*	B	B	B	B
Organization (R)	When working, I sometimes have difficulties due to being disorganized	.778*	.017	B	B	B	B
Diligence	I often push myself very hard when trying to achieve a goal	.777*	.074*	B	B	B	B
Diligence (R)	I do only the minimum amount of work needed to get by	.794*	.004	B	B	B	B
Diligence (R)	(Procrastinates)	.777*	.005	B	B	B	B
Perfectionism	I often check my work over repeatedly to find any mistakes	.520*	.020	B	B	B	C
Perfectionism	People often call me a perfectionist	.218*	.034	C	C	B	C
Prudence (R)	I make decisions based on the feeling of the moment rather than on careful thought	.592*	.050*	B	B	B	B
Prudence (R)	I make a lot of mistakes because I don't think before I act	.802*	.025	B	B	B	B

(*continued on next page*)

Table 1 (*continued*)

Factor scale and underlying content	Item (or key theme)	Main effect η_p^2	Interaction η_p^2	Desirability function category by vignette			
				F	N	CS	G
Openness							
Aesth. appreciative (R)	I would be quite bored by a visit to an art gallery	.191*	.029	A	B	A	B
Aesth. appreciative	(Appreciates nature)	.322*	.020	B	B	B	C
Inquisitiveness	I'm interested in learning about the history and politics of other countries	.479*	.076*	B	B	B	B
Inquisitiveness	(Interested in visiting ancient ruins)	.222*	.027	A	A	A	C
Inquisitiveness (R)	(Dislikes reading encyclopedias)	.317*	.069*	A	B	A	B
Creativeness	People have often told me that I have a good imagination	.482*	.055*	B	B	B	B
Creativeness	I don't think of myself as the artistic or creative type	.089*	.042*	A	A	A	C
Unconventional (R)	I think that paying attention to radical ideas is a waste of time	.163*	.021	C	C	A	C
Unconventional	I like people who have unconventional views	.142*	.034	A	A	A	C
Unconventional (R)	I find it boring to discuss philosophy	.111*	.045*	A	A	A	B

Notes. η_p^2 = Partial eta-squared; (R) = Item of negative valence; F = Fire fighter, N = Nurse, CS = Car Sales, G = General; A = Category A (approximately flat) desirability function, B = Category B (non-linear monotonic) desirability function, C = Category C (asymmetric inverted U desirability function); Aesth. = Aesthetic.

* $p < .01$.

items bar one, indicating that the five response options generally attracted significantly different mean ratings across all four Vignette conditions. The effect sizes varied considerably with large effects being observed for many items from the Agreeableness, Conscientiousness, and Extraversion factor scale, and moderate effect sizes being typical of items from the Openness factor scale (Cohen, 1992). One-half of the items also exhibited significant interactions between Vignette and Response Option, suggesting that the different vignettes produced desirability functions that varied in some way.

3.2.1. Desirability functions

For all items (60) and Vignette groups (4), plots of the desirability functions were examined individually (240 plots in total). Generally, positively and negatively-keyed items exhibited what were essentially 'mirrored' desirability functions to one another, but item keying otherwise did not seem to systematically affect the general pattern of the functions. The noticeable exception, however, was the set of items measuring Emotionality. In most cases, the negatively-keyed Emotionality items were producing desirability functions that resembled those from positively-keyed items from the other scales (and vice versa). This simply reflects the general view that high levels of Emotionality are seen as less desirable than low levels. Nonetheless, two items did not conform consistently to this pattern. These items were "When someone I know well is unhappy, I can almost feel that person's pain myself" and "I remain unemotional even in situations where most people get very sentimental." Both of these items, positively and negatively-keyed respectively, capture content around the sentimentality facet of Emotionality, but the desirability ratings seemed to imply that high levels of Emotionality, as measured by these two items, were *more* desirable.

Broadly, the desirability functions conformed to one of three distinct categories of patterns. Desirability functions were categorized as 'approximately flat' (termed Category A) if the difference in mean ratings between the highest and lowest rated options was less than .83. This cut-off was the square root of the mean of the variances of the ratings of all response options for all items and thus represented a global estimate of within-rating variation. Fig. 1 provides a representative example of a function of this type. Of the 240 desirability functions plotted, 38 (15.8%) conformed to this pattern and these are marked in Table 1 by the letter A.

As implied by the shape of the function, items yielding desirability functions of this nature captured personality content that would plausibly be seen as largely irrelevant to the associated occupational context. For example, as expected, a high proportion of the Category A desirability functions (44.7%) emerged for items from the Openness factor scale, though none of these emerged within the general job Vignette condition. Category A desirability functions were also quite common to the items measuring Emotionality, emerging on 10 occasions, though never in the fire fighter Vignette condition.

The second category of desirability function (termed Category B) could be described as one showing a non-linear, but still monotonic, relationship between response option and mean desirability ratings. This was the most common desirability function category, emerging exactly two-thirds of the time (160 times out of a possible 240), as marked in Table 1 by the letter B. Overwhelmingly, charts in Category B were characterized by small near-linear increases in desirability across the three least desirable options, followed by a relatively large jump in desirability at the fourth option and then a relatively small increase at the fifth option. To verify this latter observation statistically, 95% confidence intervals were established around each of the mean desirability ratings within each Category B desirability function. For 139 out of the 160 Category B functions (87%), the confidence

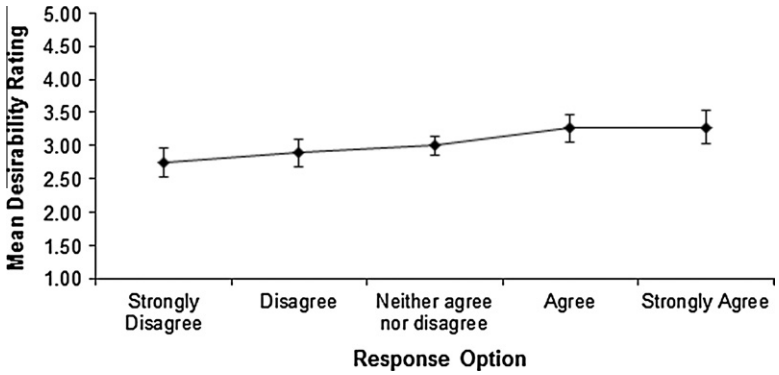

Fig. 1. Representative example of a Category A (approximately flat) desirability function with 95% confidence intervals. This item, "I like people who have unconventional views" captures content around unconventionality from within the HEXACO Openness scale. The desirability function was drawn from the Fire Fighter vignette.

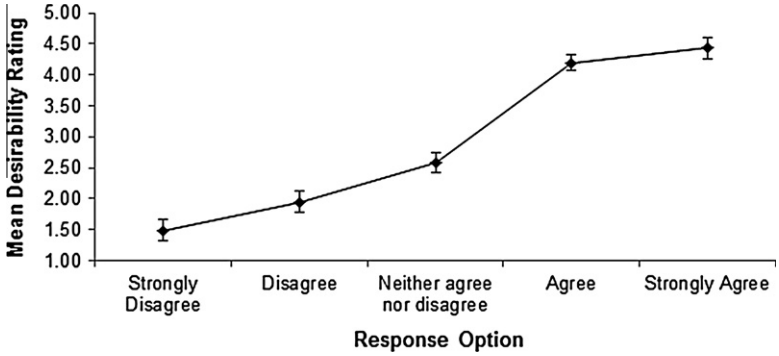

Fig. 2. Representative example of a Category B (non-linear monotonic) desirability function with 95% confidence intervals. This item, "I like to keep all my belongings stored in their proper place" captures content around organization from within the HEXACO Conscientiousness scale. The desirability function was drawn from the Nurse vignette.

intervals around the mean ratings for the two most desirable options overlapped, but the confidence intervals of the middle and fourth option did *not* overlap. Fig. 2, which provides a representative example of a Category B desirability function, also illustrates this very common pattern.

As Table 1 shows, Category B desirability functions emerged most frequently for items capturing Agreeableness (36 out of 40 possible times), followed by Conscientiousness (34 times), Honesty–Humility (30 times), and Extraversion (28 times). These functions were relatively rare for items capturing Openness (16 times) and Emotionality (15 times).

The third category, asymmetric inverted U (Category C) was very similar to the Category B but with a *decrease* in desirability at the fifth option relative to the fourth, thus producing a 'bend' at the fourth rating option. Fig. 3 provides a representative example of a function of this type. A total of 42 desirability functions (17.5%) conformed to such a pattern (marked in Table 1 with the letter C), though for the great majority of these, the decline in mean desirability at the fifth point was only very slight. Indeed, for all Category C desirability functions, the 95% confidence intervals around the mean desirability ratings for the fourth and fifth options overlapped. We acknowledge that the overlapping confidence intervals suggest that Category C functions are arguably *statistically* similar to Category B functions. Nonetheless, we elected to assign these functions into a separate category because they were consistent with the most dominant pattern observed by Kuncel and Tellegen (2009) and they represent a functionally distinct interpretation of the desirability of extreme levels of traits.

Items measuring the Emotionality factor exhibited the most instances of Category C functions (15 out of a possible 40 times), though these functions also appeared with some regularity in the items measuring Extraversion (8 items) and Openness (7 times). Many items exhibiting Category A functions in some Vignette conditions were also exhibiting Category C functions in others. In the rare instances where this type of desirability function emerged in at least three of the four conditions, the item seemed to contain quite strong language or wording, perhaps capturing extreme levels of the underlying trait (e.g. "I think paying attention to *radical* ideas is a *waste of time*", "People often regard me as a *perfectionist*", "When someone I know well is unhappy, I can almost *feel that person's pain myself*").

3.2.2. Impact of context

Whilst some contextual effects were touched on above, more formal tests of the moderating impact of Vignette are provided by examining the interaction terms. Indeed, a significant interaction between Vignette and Response Option was observed for 30 of the 60 items (see Table 1). In almost all cases, the sizes of the interaction effects were somewhat smaller than those of the main effects, and they all fell within the small to medium range as described by Cohen (1992).

The next stage of the analysis was to plot the desirability functions of these 30 items for all Vignettes together so as to ascertain the nature of the interactions. In many cases, the significant interaction term could be clearly attributed to the desirability function for the car salesperson vignette being quite different to those of the remaining three. As expected, such items included those measuring Honesty–Humility and Extraversion, plus one item capturing content around flexibility from within the Agreeableness factor. Figs. 4–6 provide some representative examples of the observed patterns.

Significant interaction terms were also observed for the items that capture content around fearfulness within the Emotionality factor. Inspection of the resultant means plots for the first two

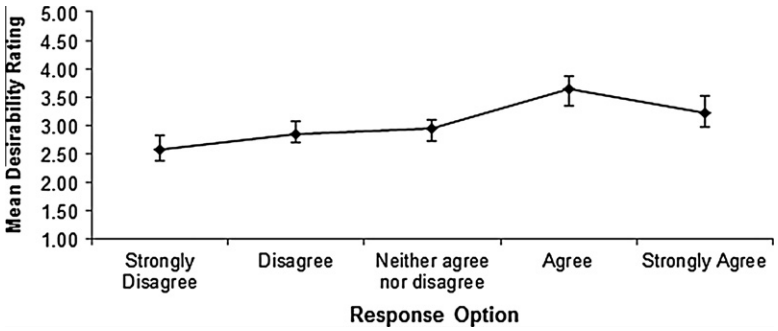

Fig. 3. Representative example of a Category C (asymmetric inverted U) desirability function with 95% confidence intervals. This item, "People often call me a perfectionist" captures content around perfectionism from within the HEXACO Conscientiousness scale. The desirability function was drawn from the General Job vignette.

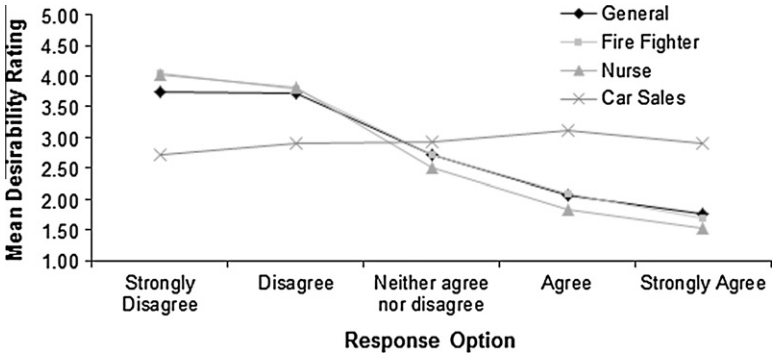

Fig. 4. Desirability functions for all Vignettes from an Honesty–Humility item which revealed a different pattern for the Car Sales vignette. Endorsement of this item, which refers to taking risks to improve one's social status, reflects low levels of greed avoidance.

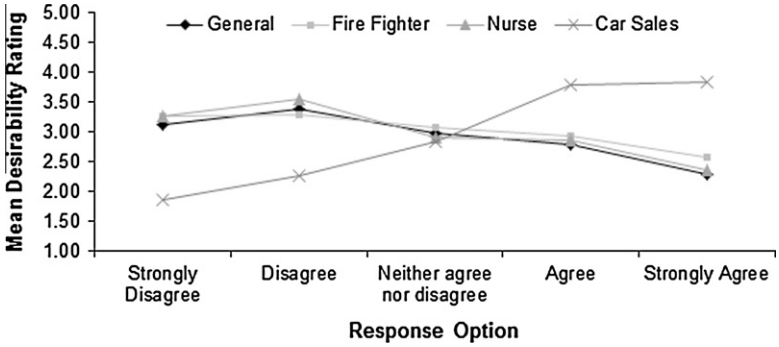

Fig. 5. Desirability functions for all Vignettes from an Extraversion item which revealed a different pattern for the Car Sales vignette. Endorsement of this item, which refers to doing more talking than others when engaged in conversation, represents high levels of expressiveness.

items, presented in Figs. 7 and 8, revealed that the significant interactions could be attributed to the fire fighter *and* nurse vignettes. Interestingly, for the nurse vignette, the two desirability functions are in direct conflict with one another, despite the items capturing similar content within the Emotionality factor scale. The desirability function of the first item suggests that high levels of fearfulness are *more* desirable for a nurse (relative to the other occupations) whereas the desirability function of the second item implies the opposite.

Only one of the items capturing content around anxiety in the Emotionality factor scale exhibited a significant interaction, plotted in Fig. 9, and for this item the differences in desirability functions across the vignettes were nowhere near as pronounced as they were for items capturing content around fearfulness. Further, only one of the items measuring traits that were expected to be uniquely relevant to the nurse condition exhibited a significant interaction. A means plot for this item, shown in Fig. 10, was difficult to interpret and certainly not suggestive that the desirability function for the nurse vignette was that which prompted the significant interaction, and indeed, the interaction effect size was very small $\left(\eta_p^2 = .046\right)$. Nonetheless, the results do support the general hypothesis that the occupational context impacts upon the desirability functions.

4. Discussion

4.1. Summary

The present study revealed three important findings. First, under conditions where participants were presented with personality items and asked to rate the desirability of the response

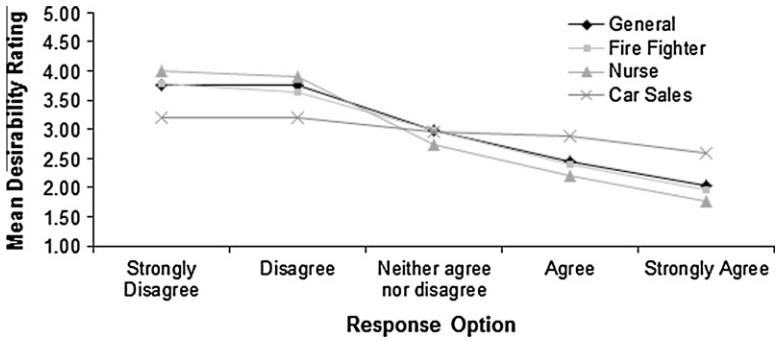

Fig. 6. Desirability functions for all Vignettes from an Agreeableness item which revealed a different pattern for the Car Sales vignette. Endorsement of this item, "People sometimes tell me that I'm too stubborn", reflects low levels of flexibility.

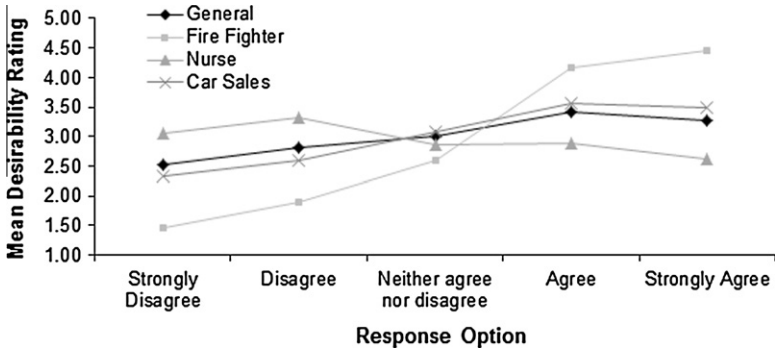

Fig. 7. Desirability functions for all Vignettes from an Emotionality item which refers to resilience to physical pain. Endorsement of this item reflects lower levels of fearfulness.

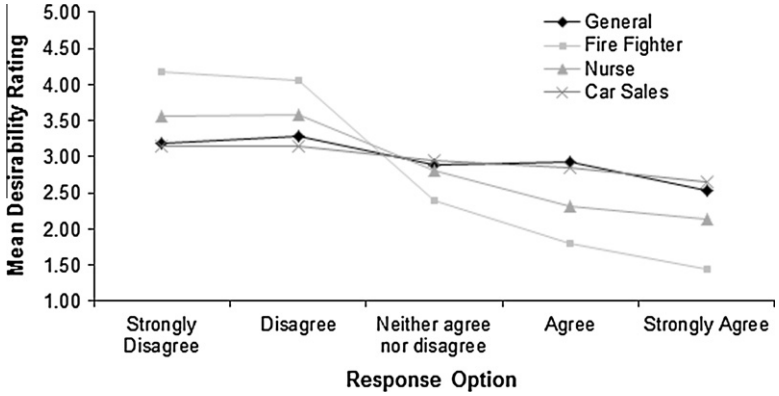

Fig. 8. Desirability functions for all Vignettes from the Emotionality item, "When it comes to physical danger, I am very fearful". Endorsement of this item reflects higher levels of fearfulness.

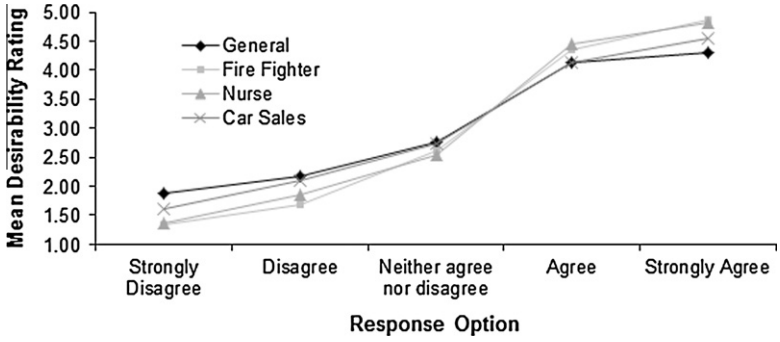

Fig. 9. Desirability functions for all Vignettes from an Emotionality item which refers to remaining calm. Endorsement of this item reflects lower levels of anxiety.

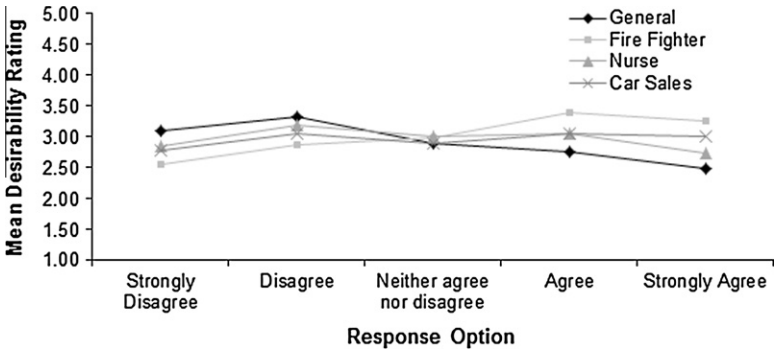

Fig. 10. Desirability functions for all Vignettes from the Emotionality item, "I remain unemotional even in situations where most people get very sentimental". Endorsement of this item reflects higher levels of sentimentality.

options Strongly Disagree through to Strongly Agree, the resultant desirability functions that emerged were non-linear, though in contrast to previous research, still mostly monotonic. Second, the perceptions of the desirability of responses to many personality items depended on the context of the target job. Finally, in some cases, desirability functions even varied across items intended to capture the same content within a personality domain. These three points are discussed in more depth below.

4.2. Comparisons to previous research

In the present study, the personality items generally failed to demonstrate clear evidence of a linear relationship between perceived desirability and the levels of agreement with personality descriptive statements, a result conceptually similar to those observed by Borkenau et al. (2009) and Kuncel and Tellegen (2009). Nonetheless, the general patterns of results observed in this study were quite different to those of Kuncel and Tellegen's first study, despite using a similar method. Indeed, very few items in the present study yielded desirability functions incorporating an inverted U shape, and none of the functions that did incorporate an inverted U were quasi-symmetric. Even for the functions that did include an inverted U shape, the difference in desirability between the most extreme and penultimate response options tended to be small. Instead, the most dominant pattern of results suggested that the psychological difference in perceived desirability between Neither and the next most desirable option (i.e. Agree for positively-keyed items and Disagree for negatively-keyed items) was much larger than that between any other pair of adjacent response options.

The prime suspects for the disparity in results between this study and those of Kuncel and Tellegen's (2009) study are the full self-descriptive statements used to measure personality (as opposed to adjectival trait measures) combined with the subjective Likert-type scale used to gauge trait levels (as opposed to specific points on a trait continuum). The asymmetric inverted U functions observed most often by Kuncel and Tellegen imply that extreme levels of many personality traits are viewed as undesirable. In that study, participants were informed, first, exactly *what* trait was being considered (as per the adjective presented to them) and, second, exactly what *level* of that trait was being considered. In the present study, however, participants are left to their own judgments about what trait a particular personality item is measuring and the levels of that trait being captured by the response scale. For example, participants in this study might judge the statement "I prefer jobs that involve active social interaction to those that involve working alone" as reflecting a typical person's level of Extraversion, whereas they might judge "People often call me a perfectionist" as reflecting an extreme level of Conscientiousness. By association, the Strongly Agree response option, whilst the most extreme option available, possibly reflects quite different levels of the Extraversion and Conscientiousness traits respectively (whereas a response option reading "extremely high – top 1%" does not). In support of this notion, the few items that exhibited desirability functions incorporating an inverted U shape across multiple vignettes seemed to incorporate some quite extreme wording. The results of this study, when contrasted against those of Kuncel and Tellegen, therefore suggest that trait *levels* themselves may not be linearly related to the Strongly Disagree—Strongly Agree response options, and/or that the full spectrum of a trait is not encapsulated by a Strongly Disagree—Strongly Agree response scale. To investigate this possibility would require the use of advanced item analyses techniques such as item-response theory, using partial-credit models. Further, some recent research has suggested that measurement via relative referencing (i.e. asking people to consider where they sit on a trait, relative to others) may also provide an avenue from which to better capture extreme levels of traits (see Goffin, Jelley, Powell, & Johnston, 2009; Goffin & Olson, 2011).

Another possible cause of the disparity in results between Kuncel and Tellegen's (2009) study and this one is the present study's context of a hypothetical person applying for a job. Whilst Kuncel and Tellegen did ask one group of participants to consider the desirability of different levels of personality traits at work, that instruction is subtly different to asking participants to consider the desirability of different levels of traits, at least as they are captured by a Likert-type scale, when *applying for a job*. For example, participants may have interpreted the Neither option as synonymous to being average. Organizations do not, however, typically express an overt desire to employ 'average' individuals, and applicants are similarly unlikely to want to portray themselves as being 'average', even if average levels of traits would be most seen by others as most desirable *once somebody is actually in the job*. The hypothetical job application context may have therefore lowered the perceived desirability of the middle response category, and hence exaggerated the differences between this option and the next most desirable options. It may have also explained why no asymmetric inverted-U desirability functions (i.e. where being average on a trait is most desirable) were observed in this study (c.f. Kuncel & Tellegen, 2009). Perhaps if an 'everyday' (i.e. non-applicant) context had been included in this study, the Neither option would have attracted more favorable ratings.

4.3. Occupational context

The presence of significant interaction terms for half of the items suggests that the responses deemed most desirable to applicants vary depending on the nature of the target job. This result is largely congruent with previous findings which have demonstrated

that individuals, when instructed to fake, are able to vary their strategy according to the nature of the target context (e.g. Dalen, Stanton, & Roberts, 2001; Krahe, 1989; Pauls & Crost, 2005). Even for those items for which interactions were not observed in the present study, there still may be scope for contextual effects to emerge given the right vignette. For example, one could imagine that a job of artist or professor might attract differential desirability functions for the Openness items.

There may, however, be certain items that produce desirability functions that are largely context-resistant or, in other words, traits for which the value attributed to different levels generalizes across many or all jobs. Some clues as to the types of traits that might meet this criterion are provided in the present study amongst the Agreeableness items where only one of the items from this scale exhibited a significant interaction.

4.4. Implications for research into personality and socially desirable responding

In spite of the high proportion of desirability functions revealing monotonic relationships between response option and desirability, we argue here that an individual wishing to respond in a socially desirable manner may not necessarily select the most extreme response when faced with an item measuring a desirable personality trait. This conclusion might initially seem at odds with the observation that the majority of items did evidence higher desirability ratings for the extreme response. The critical element here lies, however, in the very small difference between the extreme and penultimate responses observed for almost all the items. Given the nature of the pattern of desirability function that emerged with most consistency, for most personality items, it seems that an individual in a high-stakes setting (e.g. job applicant) who wishes to be portrayed in a desirable manner may not feel it matters much for the overall desirability of the profile whether he or she Agrees or Strongly Agrees with a positively-keyed personality item. There are, however, other reasons why extreme responses might be less appealing in high-stakes settings. For example, a job applicant might feel that the more extreme response carries with it a greater risk of being caught out as a 'faker', or that it would be difficult to emulate the behavior reflected by the extreme response (Goffin & Boyd, 2009; Kuncel & Tellegen, 2009). This extreme option does not, however, appear to offer much additional value in the way of perceived desirability.

Further, in most cases, responding Neither to a desirable item appeared to be seen as being *considerably* less desirable than responding Agree. Perhaps, then, individuals who are trying to manipulate their responses so as to be portrayed favorably might be best characterized by an apparent reluctance to choose the much less desirable middle or lower options rather than a tendency to preferentially select the extreme options. Much of the research into the socially desirable responding phenomenon has simply compared mean personality scale scores of fakers and non-fakers, or job applicants to incumbents or research participants (e.g. Birkeland et al., 2006). As an alternative avenue of research, it seems likely to be worthwhile drawing from techniques, such as item response models, to explore the *patterns* of responses from individuals in high-stakes settings, such as job applicants (e.g. Griffin et al., 2004). By doing so, it becomes possible to gain a clearer understanding of the psychological distinctions made between response options items in high-stakes settings and how item wording might affect these. As it stands, classical testing models treat all items from the same scale as being equivalent, whereas the present study suggests that this might not be appropriate.

Indeed, another telling finding from this study was that desirability functions varied, in some cases quite considerably, across items from the same personality factor scale designed to assess similar content. Such a result reinforces the notion that focusing exclusively on individuals' personality scale scores in studies of impression management will mask some of the complexities that underpin socially desirable response patterns. Only relatively few studies have explored faking through at least some analyses at the individual item level (e.g. Griffin et al., 2004; Kuncel & Borneman, 2007) but such work has suggested that some items do indeed 'behave' differently to others in the same scale under high impression management conditions.

4.5. Limitations

The findings of the present study should be tempered with acknowledgment of its inherent methodological limitations. By recruiting non-applicants for this study, we may be introducing potential biases to the extent that individuals who apply for the jobs represented by the vignettes regard the desirability of different traits differently to others. To investigate the issue further, future research might compare the desirability functions produced by applicants or incumbents to those produced by non-applicants. While it seems reasonable that applicants or incumbents of the various jobs explored here might provide different sets of desirability functions to non-applicants in relation to their specific roles, we very much doubt that the general conclusions of this study – that the occupational context moderates the perceived desirability of responses to items – would be different.

It could also be argued that the vignettes used in this study were not sufficiently representative of the information applicants typically receive when completing a personality questionnaire for a job. Insight into how the participants might have used the information they received to make judgments on desirability is offered by Dalen et al. (2001), who found that people tend to fake to a profile that is based on their own stereotypical view of an incumbent (see also Furnham, 1990; Mahar et al., 1995). Tellingly, Dalen et al. found that the same stereotypes emerged *irrespective of the amount of information the participants received about the target job*. Though in the present study, the vignettes were reasonably detailed, one might speculate that the contextual effects would have emerged to a similar extent if participants were provided with job titles only.

4.6. Concluding statements

The results of the present study provide further evidence of non-linear relationships between trait level and perceptions of desirability. Findings such as these and those of Kuncel and Tellegen (2009), and Borkenau et al. (2009) suggest a need to call into question the validity of simply examining mean scale scores on personality factors in future studies of impression management. For example, in detecting fakers it may be a worthwhile avenue to compare responses to a predefined set of 'stereotypic' responses as a screening tool (Dalen et al., 2001). The present findings also suggest that idiosyncratic socially desirable responding may be a phenomenon best examined at the item rather than scale level. We therefore feel that it is time for researchers to apply more advanced response analysis techniques when investigating the phenomenon of socially desirable responding.

Appendix A. Job application vignettes

A.1. Vignette for the fire-fighter condition

"J" has applied to be a *Fire Fighter.* As a Fire Fighter, "J" would work as a member of a cohesive and cooperative team, under supervision, whether involved in emergency response or more

routine work associated with maintaining readiness or community safety activities. "J" must be physically fit and able to work in extreme conditions, often exposed to the elements. "J" would also be required to act as a representative of the emergency services organization to members of local communities. "J" will be assessed against the following selection criteria:

- demonstrated ability to work as an effective member of a highly trained team,
- demonstrated ability to solve problems in a stressful and physically demanding environment,
- demonstrated interpersonal and communication skills,
- demonstrated commitment to excellence in job performance and ongoing personal and professional development,
- demonstrated ability to work successfully with, and relate to a diverse range of people, and
- demonstrated integrity.

A.2. Vignette for the nurse condition

"J" has applied to be a *Nurse*. As a Nurse, "J" would be required to work as part of a team of experienced health professionals to ensure that the hospital operates smoothly. "J" must ensure that the patients are looked after and are as comfortable as possible within the hospital environment. "J" must co-ordinate and assist in a wide variety of treatments and care activities for patients. "J" will be assessed against the following selection criteria:

- A caring and compassionate attitude towards patients, whilst still remaining professional.
- Excellent communication and interpersonal skills with colleagues and patients.
- Sound knowledge of hospital procedures.
- Sound computer skills.
- Demonstrated ability to build rapport quickly and relate to a diverse range of people.

A.3. Vignette for the used car salesperson condition

"J" has applied to be a *Used Car Salesperson.* As a *Used Car Salesperson,* "J" would be required to meet with customers on a daily basis to discuss the types of vehicle they are interested in purchasing. "J" would then find customers a vehicle which fits their criteria. "J" would need to have a sound understanding of all of the vehicles for sale. "J" would also be approachable and well presented and comfortable negotiating with customers. "J" will be assessed against the following selection criteria:

- strong interpersonal and communication skills to facilitate relationships with customers,
- strong negotiating and influencing skills,
- demonstrated knowledge of all vehicles for sale,
- an energetic, friendly and enthusiastic demeanor, and
- a strong desire for success, both personally and for the customer.

Appendix B. Instructions for participants

Many organizations now use personality assessments as part of their application process for prospective employees. In this exercise, imagine that a person called "J" has applied for a job [as a fire fighter/nurse/car sales person] and has been asked to complete a personality questionnaire as part of the application. You will be presented with the set of behavioral statements that "J" will have to respond to when taking the personality questionnaire. Underneath each statement are the five different possible responses that "J" can provide to that statement. Your task is to indicate, for each of the five possible responses, how desirable "J" would appear to a hiring organization if he or she selected that option. To do so, please use the rating scale which is provided.

References

Ashton, M. C., & Lee, K. (2001). A theoretical basis for the major dimensions of personality. *European Journal of Personality, 15*, 327–353.

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*, 150–166.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*, 317–335.

Borkenau, P., Zaltauskas, K., & Leising, D. (2009). More may be better but there may be too much: Optimal trait level and self-enhancement bias. *Journal of Personality, 77*, 825–858.

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

Dalen, L. H., Stanton, N. A., & Roberts, A. D. (2001). Faking personality questionnaires in personnel selection. *The Journal of Management Development, 20*, 729–742.

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology, 37*, 90–93.

Furnham, A. (1990). Faking personality questionnaires: Fabricating different profiles for different purposes. *Current Psychology: Research & Reviews, 9*, 46–55.

Goffin, R. D., & Boyd, A. C. (2009). Faking and personality assessment in personnel selection: Advancing models of faking. *Canadian Psychology, 50*, 151–160.

Goffin, R. D., Jelley, R., Powell, D. M., & Johnston, N. G. (2009). Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management, 48*, 251–268.

Goffin, R. D., & Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science, 6*, 48–60.

Griffin, B., Hesketh, B., & Grayson, D. (2004). Applicants faking good: Evidence of item bias in the NEO PI-R. *Personality and Individual Differences, 36*, 1545–1558.

Krahe, B. (1989). Faking personality profiles on a standard personality inventory. *Personality and Individual Differences, 10*, 437–443.

Kuncel, N. R., & Borneman, M. J. (2007). Toward a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment, 15*, 220–231.

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology, 62*, 201–228.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research, 39*(2), 329–358.

Lee, K., & Ashton, M. C. (2005). Psychopathy, machiavellianism, and narcissism in the five-factor model and the HEXACO model of personality structure. *Personality and Individual Differences, 38*, 1571–1582.

Mahar, D., Coburn, B., Griffin, N., Hemeter, F., Potappel, C., Turton, M., et al. (2006). Stereotyping as a response strategy when faking personality questionnaires. *Personality and Individual Differences, 40*, 1375–1386.

Mahar, D., Cologon, J., & Duck, J. (1995). Response strategies when faking personality questionnaires in a vocational selection setting. *Personality and Individual Differences, 18*, 605–609.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683–729.

Pauls, C. A., & Crost, N. W. (2005). Effects of different instructional sets on the construct validity of the NEO-PI-R. *Personality and Individual Differences, 39*, 297–308.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155–180.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9*, 219–242.

Yorke, M. (2001). Bipolarity. . .or not? Some conceptual problems relating to bipolar rating scales. *British Educational Research Journal, 27*, 171–186.