The Sensitivity of the CPI Good Impression Scale for Detecting ‘Faking Good’ among Norwegian Students and Job Applicants

Gro M. Sandal* and Inger M. Endresen
University of Bergen

This study assessed the sensitivity of the California Psychological Inventory’s Good Impression scale in the detection of ‘faking good’ in Norway. The original development of the scale was replicated using Norwegian students (N = 248). Paired sample T-tests revealed significant differences for 31 of the 40 items when scores from normal test conditions were contrasted with scores obtained under ‘fake good’ instructions. Further, job applicants (N = 494) scored significantly higher on the Good Impression scale than students. Comparisons of mean scale scores between Norwegian and US student samples suggested that the Norwegian students displayed a more modest self-presentation. The Commonality Scale of the CPI was employed to investigate differences in scale usage between these two cultures. The analyses suggested that the differences in the Good Impression scale observed between these two cultures could not be explained by differences in acquiescence tendencies. Results suggest that the Good Impression scale is able to capture response distortion in Norwegian personnel selection situations, but modification of some items is needed.

Introduction

In the use of personality scales for the assessment of job applicants in applied settings, the impact of social desirability is often cited as a concern. Given the seemingly straightforward or transparent nature of some items in such personality measures, it seems likely that a number of applicants will distort their responses and try to ‘beat the test’ (Viswesvaran and Ones 1999). To manage the problem of response distortion, several personality inventories include validity scales used to identify candidates who attempt to present themselves in an excessively favourable light. Impression-making, or ‘faking good’ scales typically include items that have been developed to detect a conscious form of self-presentation motivated by the desire to obtain social approval. However, little attention has been paid to the cross-cultural applicability of such scales even though it is well documented that response sets may differ across societies (e.g., Heine and Lehman 1995; Hofstede 1980; Mwamwenda 1996; Smith, Smith and Seymour 1993; Warnecke, Johnson, Chavez, Sudman, O’Rourke, Lacey and Horn 1997). In multinational pools of applicants, cultural fairness in employment decisions is of particular concern. Furthermore, the cross-cultural applicability of such tests is an important consideration given the widespread use of translated tests in many countries.

Assessment of Social Desirable Test Responses

One aim of the present article is to evaluate the ability of the Good Impression scale of the California Psychological Inventory (CPI; Gough 1952, 1987; Gough and Bradley 1996) to detect tendencies of ‘faking good’ among Norwegian job applicants. Given that the scales in the CPI were designed to represent ‘dispositions having universal status’ thought to be relevant to any and all cultures, the inventory is particularly appropriate for cross-cultural validation. Many studies have been reported describing the adoption and use of the CPI in diverse cultures (Paunonen and Ashton 1997). We have found no studies, however, that have reported psychometric data of the kind necessary for evaluating
the invariance of the Good Impression scale’s properties across cultures. As for most CPI scales, the original Good Impression scale was constructed using an empirical approach where items were selected according to their relation to empirical criteria. In the development of the original scale, subjects first completed an experimental booklet of items under normal circumstances. Following this, the testing was repeated with instructions to present a very favourable self-portrait — the kind they might give if applying for a very much desired job, for example, or in any situations in which they wanted to be judged as admirable and praiseworthy. By contrasting the results of these two sets of protocols, items were identified that yielded significant change in degrees of endorsement. This analysis gave rise to the original 40-item Good Impression scale of the CPI, on which a very high score is believed to be indicative of ‘faking good’. In the present study, the original development of the Good Impression scale was replicated in a sample of Norwegian students. It has been questioned whether results from faking studies generalize to real-world application settings (Goffin and Woods 1995; Hough and Schneider 1996; Ones and Viswesvaran 1998). Therefore, the scores of the students under the two conditions were compared with those of applicants to managerial positions.

The second aim of this article is to examine potential cross-cultural differences between Norway and the United States in scores on the Good Impression scale. Several studies have demonstrated that most people in Western societies perceive themselves as being above average on a wide variety of desirable traits (see Kurman and Sriram 1997). In North American samples, a strong and positive correlation has been found between item desirability and endorsement (Smith et al. 1993). Whether this relation replicates cross-culturally in other Western societies has received less attention. In the Norwegian culture, normative pressures against self-promotion may contribute to a less positive self-description than what is normally found among North American respondents. In an extensive cross-cultural study by Hofstede (1980), Norway was found to score very low on Masculinity – a dimension that distinguishes between cultures in which assertiveness and ambition are highly valued (so-called masculine cultures) and cultures in which greater emphasis is placed on cooperation and good working relationships (so-called feminine cultures). In comparison to the United States, Norway also scored lower on Hofstede’s dimension of Power Distance. This dimension refers to the extent to which people accept and expect that power be unequally distributed. In the Norwegian culture, these values may be reflected in such phrases as ‘be who you are’ or, ‘you ought not to believe that you are better than us’. In American eyes, Norwegian job applicants might tend to undersell themselves in that they are careful not to make promises they are not certain they can fulfil.

Gender Roles and Social Desirability

The content of social gender roles represents one potential source of variance across cultures in terms of social desirability (e.g., Watkins and Cheung 1995), and the present study therefore also examined main and interaction effects of gender and culture. Normative data for US students has shown no significant gender difference on the Good Impression scale (Gough and Bradley 1996), but some researchers have questioned whether students give the same responses as organizational members (cf. Gordon, Slade and Schmitt 1986). Critics have argued that the practice of adjusting substantive personality scale scores on the basis of social desirability scales might cause adverse impacts for female applicants because items often refer to stereotypically feminine characteristics, such as selflessness, other-directness and nurturing (e.g., Gannon, Raber, Jenkins, Kettermann and Griffith 1997). Despite this assumption, results from national and cross-cultural studies on gender differences in social desirability responding have been inconsistent (Ones and Viswesvaran 1998). Contrasting results are likely explained by variation in the extent to which social gender roles are clearly distinct. In so-called feminine societies, such as Norway, overlapping gender roles might serve to minimize differences in social desirability responding since both men and women are expected to be modest.

Hypotheses

Previous research has found a high cross-cultural rate of agreement on which traits are desirable or undesirable (see Smith et al. 1993). In the present study it was therefore hypothesized that most items in the Good Impression scale would discriminate between the normal and ‘fake good’ conditions also among Norwegian students (hypothesis 1). Due to a strong emphasis on equality and modesty in Norway, it was expected that the scores on the Good Impression scale for the Norwegian job applicants in a real-life setting would closely resemble those of students under normal test instructions (hypothesis 2). Based on the assumption that Norwegian culture emphasizes modesty and is less conducive to competitive self-promotion than mainstream North American culture, it was hypothesized that Norwegian students would score lower than their US counterparts on the Good Impression scale (hypothesis 3). No gender differences in the Norwegian samples were expected either among students or among job applicants (hypothesis 4).

One methodological bias in conducting cross-cultural research in personality assessment is that demonstrated differences may be due to differences in scale usage. Several studies have documented variations across cultures in tendencies to acquiesce, or agree with items independent of their content (see Grimm and Church
Available evidence on potential differences between Western European and US cultures are sparse and more research is clearly needed. As an attempt to rule out this factor as an alternative explanation for potential mean differences on the Good Impression scale between Norwegian and US samples, items from a second CPI scale, called Commonality, were administered together with the Good Impression items. Commonality was developed to identify protocols rendered invalid by careless, random, unknowing or extremely unusual patterns of response to the inventory. Half of the items in the scale are answered ‘true’ by a very high percentage of US respondents, constituting one half of the scale (e.g., ‘there are a few people who just cannot be trusted’) buttressed by the other half of the scale composed of items seldom if ever answered ‘true’ (e.g., ‘I cannot do anything well’). Potential cross-cultural differences in acquiescence might be detected in a higher rate of ‘true’ responses to the latter subset of items. In line with the conception of Norwegians as being modest and cautious in their self-descriptions, it was hypothesized that Norwegian students would not acquiesce more than their US counterparts (hypothesis 5).

Method

Sample

A total of 248 first-year students enrolled in introductory courses at the University of Bergen took part in the study. The average age of the sample was 22.2 years (SD = 8.5). Of the 494 applicants for managerial positions, with an average age of 38.7 years (SD = 8.5). Of the 494 managerial applicants, 21.5% were women, 58.7% already held position on a superior management level whereas 41.3% were mid-level managers.

Instruments

The Form 434 of the California Psychological Inventory (Gough and Bradley 1996) is a 434-item questionnaire that requires a minimum of fourth grade reading ability and can typically be completed within 45 to 60 minutes. Each item consists of a single sentence about how the person generally feels or has been treated, and to which the responder answers either true (0) or false (1), depending on whether the item is found to be descriptive or non-descriptive. Originally, the instrument had 480 items (Gough 1952). In the two revised versions, Form 462 (Gough 1987) and Form 434, 18 items have been dropped (12 of which were ‘repeated items’ that occurred twice in the inventory), and 29 other items have been changed ‘so as to make them easier to understand, more up-to-date, and less encrusted with sexist or undesirable phraseology’ (Gough and Bradley 1996). The items are combined to produce a total of 20 scales assumed to measure ‘folk concepts’, or in other words, concepts that arise from and are linked to the processes of interpersonal life, and that are to be found everywhere that humans congregate into groups or societal functions. Among the 20 scales are three validation scales constructed for the detection of invalid protocols. Two of them, the Good Impression and the Commonality scales are part of the present study.

The 40-item Good Impression (GI) scale contains items showing significant differences in endorsement under normal and ‘fake good’ conditions (Gough 1952; Gough and Bradley 1996). Before calculating scale scores, responses to 10 of the items are reversed. High scale scores are interpreted as wanting to make a good impression whereas low scores suggest more realistic self-representation, even if this causes frictions or problems. Values of 30 and above are considered suggestive of ‘faking good’ (Gough 1987).

The 38-item Commonality (Cm) scale includes 38 items to which a very large percentage of respondents have been found to give the same response. Before scoring the Commonality scale, responses to the items that conventionally are answered ‘false’ are reversed. High scores on the Commonality scale suggest that the respondent has answered the inventory in a standard or modal fashion, whereas very low scores (raw scores of 24 and below) raise the possibility of erratic, random or strongly non-normative responses. Apart from detecting invalid protocols, the Commonality scale is also assumed to carry interpretative implications for normal protocols. Yet another function of this scale is to assess a continuum going from conventionality at the high end to non-conventionality at the low end. Moderately low scores (30 to 34) are taken as indicative of independence of mind, and creative differences from the norm, while high scores (raw scores of 37 to 38) are assumed to indicate, aside from validity, conventionality and conformity.

Procedure

Before answering the questionnaires, students signed a Declaration of Consent stating that their participation was voluntary and that data would be stored and treated in a confidential manner. The test booklet administered to the students contained a total of 98 items from the Good Impression, Commonality and FM (Femininity/ Masculinity) scale in a randomized order. Items from the last scale were included to prevent the research questions from being transparent to the respondents and these results are not reported in this article. The questionnaire was completed first under normal (‘straight-take’), and second, under ‘fake good’ instructions with no interruption between the two test administrations. Under the normal conditions, subjects were simply instructed to...
answer the questions. Under the ‘fake good’ conditions, the subjects were instructed to answer the questions as if they were applying for a desirable position. Instructions were read aloud by a research assistant in addition to being provided on a cover sheet. At the end of the testing session, subjects were debriefed as to the nature of the study. The job applicants received the standardized administration of the CPI as part of the employment application process. Current employment status, age and gender were indicated on a cover sheet. All applicants received feedback on their own results as part of a subsequent job interview.

Aggregated scores on the Good Impression scale and the Commonality scale were formed by computing the mean of the responses on the individual items for each scale, thereafter multiplying these with the total number of items. This procedure was chosen (instead of computing sum scores) since some of the subjects had missing data for individual items. Data from US samples were taken from the CPI manual Form 434 (Gough and Bradley 1996). Results for US student sample were based on college students, general sample (males \( n = 3.235 \), ibid., p. 183 and females \( n = 4.126 \), ibid., p. 190). The students’ scores in the ‘fake good’ condition were compared with simulated protocols for US males (\( n = 50 \), ibid, p. 188) and US females (\( n = 52 \), ibid., p. 193). In comparing the average percentage of true scores on the Commonality scale in the Norwegian sample versus the basic normative US sample, the Itemmetric data reported in the manual were used (ibid., pp. 400–409).

**Results**

**Descriptive Statistics and Internal Consistency for the Good Impression Scale**

Table 1 shows the mean values for the Good Impression scale for the total sample, the group of job applicants, and students. The internal consistency for the total sample, measured by Cronbach’s alpha was .83. Cronbach’s alphas for the two sub-samples (students and job applicants) and separated by gender are also shown in Table 1.

The frequency distribution of the raw scores on the Good Impression scale was examined separately for students under the two test instructions and for the job applicants. Raw scores on the Good Impression scale of 30 and above are interpreted as suggestive of ‘faking good’. Only .5% of the students scored in this area under normal test instructions compared to 13.6% when they were instructed to fake good. The corresponding result for job applicants was 13%, which is almost identical as for students under ‘faking good’ instructions. Raw scores on the Good Impression scale from 18 to 29 are assumed to reflect an increasing concern about the reactions of others (Gough and Bradley 1996). Under normal test instructions, 23.6% of the students scored within this range. The corresponding results for students under ‘faking good’ test instructions and job applicants were 56.5% and 71.8%, respectively.

**Gender and Age Differences**

T-test for independent samples were used to examine possible gender differences among students on the Good Impression scale under both normal and fake good test instructions. None of these differences were significant (\( t (246) = .01, p = .99; t (243) = 1.68, p = .094 \), respectively). Neither was there any significant gender effect among the group of job applicants (normal test instructions, \( t (492) = 1.15, p = .25 \)).

There was a significant age difference between the two groups (\( t (610) = 21.86, p < .001 \)). Results from correlation analyses between age and GI-scores in the two groups showed a significant positive association in the student sample (\( r = .19, p < .05 \)), while the correlation was non-significant in the group of job applicants.

**Differences between Normal and ‘Fake Good’ Test Instructions**

Paired t-tests were performed to examine if the students’ scores on the Good Impression obtained under ‘fake good’ instructions differed significantly from scores from normal test conditions. The result showed that the students scored significantly higher under ‘fake good’ instructions (\( t (245) = 15.3, p < .001 \)). The results from paired t-tests for the individual items showed significant differences in predicted direction for most items. Nine items did not discriminate between the two conditions.

**Comparison between Students and Job Applicants**

Possible differences between students (‘normal condition’) and job applicants in relation to scores on the Good Impression scale were examined by analysis of variance with group and gender as independent variables and age as covariate. The gender variable was included to examine possible interaction effect for group by gender. The analysis yielded highly significant group effect (see Table 2), while the gender and age effects were non-significant. Also effect size was calculated using Cohen’s d (\( d = (X1−X2)/s \)), where X1 and X2 are mean values on Good Impression for the two groups, and s is the pooled standard deviation; (see Cohen 1977). This effect was quite large, 1.8 (values above .8 are considered as a large effect, according to Cohen). The variable of gender did not give a significant contribution in this analysis.

Analysis of variance were also performed to examine differences between the scores of students under normal
test instructions and those of job applicants for the individual items in the Good Impressions scale. Most of these analyses yielded significant results (job applicants had the highest mean scores), but again the differences for nine of the items were non-significant. Four of these items overlapped with those items not discriminating between the student scores in the normal vs ‘fake good’ condition.

We further compared the students’ scores on the Good Impression scale when ‘fake good’ instructions were given with the job applicants’ scores (normal condition). Even if the scores in the group of job applicants were somewhat higher, the group effect was not significant when controlling for gender and age (see Table 2). The effect size for this difference was also rather low (.34).

**Comparisons between Norwegian and US Samples**

T-tests for independent samples were used to examine possible differences between Norwegian and US samples.

For all analyses, the US respondents scored significantly higher than the Norwegian respondents (see Figure 1). Contrasting the scores of the Norwegian students (normal condition, see Figure 1) with scores reported for US college students yielded significant effects both for females ($t(4293) = 7.7, p < .001$, two tailed) and males ($t(3312) = 4.1, p < .001$, two tailed). Comparisons were also made between the students’ scores in the ‘fake good’ condition and scores from simulated US protocols. Again the group differences were significant for both females ($t(219) = 7.3, p < .001$, two-tailed) and males ($t(127) = 8.2, p < .001$, two-tailed).

**Examining Atypical Responses and Response Style**

The Commonality scale of the CPI was used to examine scale usage in the Norwegian samples as compared to the normative US sample. Table 3 shows the mean percentages of ‘true’ scores in the Norwegian samples (total sample, job applicants, and students), and for the

<table>
<thead>
<tr>
<th>Table 1. Mean, standard deviation and internal consistency for the Good Impression scale for the Norwegian samples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Total sample</td>
</tr>
<tr>
<td>Males</td>
</tr>
<tr>
<td>Females</td>
</tr>
<tr>
<td>Job-applicants</td>
</tr>
<tr>
<td>Males</td>
</tr>
<tr>
<td>Females</td>
</tr>
<tr>
<td>Students</td>
</tr>
<tr>
<td>Males</td>
</tr>
<tr>
<td>Females</td>
</tr>
<tr>
<td>Faking instruction</td>
</tr>
<tr>
<td>Males</td>
</tr>
<tr>
<td>Females</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Analyses of variance comparing the scores on the Good Impression scale in the student sample with the scores in the group of job applicants</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Students (normal instruction) vs Job applicants</td>
</tr>
<tr>
<td>Students (‘fake good’ instruction) vs Job applicants</td>
</tr>
</tbody>
</table>

International Journal of Selection and Assessment © Blackwell Publishers Ltd 2002
US sample for the two subsets of items. Half of the items in the Commonality scale are conventionally answered ‘true’ by a high proportion of respondents, whereas the other half are normally answered ‘false’. The mean percentage of ‘true’ answers for the former subset of items was 89% among the Norwegian respondents and about 90% for the US sample. The results for the items that are conventionally answered ‘false’ were also highly comparable, with a mean of about 4% for the Norwegian sample compared to 7% among the US respondents. The frequency distribution for the Commonality scale revealed that only one subject in the student sample had a raw score below 25, which is taken as indicative of a random or erratic response pattern. None of the job applicants scored in this area of the scale. There were also very few scores in the range 25 to 29.

Table 3. Mean percentage of ‘true’ scores in the Norwegian samples and the US normative sample for the Commonality scale

<table>
<thead>
<tr>
<th></th>
<th>Reversed items</th>
<th>Non-Reversed items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
</tr>
<tr>
<td>Total Norwegian samples</td>
<td>3.2</td>
<td>4.6</td>
</tr>
<tr>
<td>Job-applicants</td>
<td>2.6</td>
<td>2.1</td>
</tr>
<tr>
<td>Students</td>
<td>11.5</td>
<td>7.3</td>
</tr>
<tr>
<td>US normative sample</td>
<td>7.7</td>
<td>6.6</td>
</tr>
</tbody>
</table>

Notes: The non-reversed items (19 items) are conventionally answered ‘true’, whereas the reversed items are conventionally answered ‘false’. (The corresponding mean percents of ‘false’ responses are shown in parenthesis.)

Discussion

The findings from this study suggest that the Good Impression scale of the CPI is sensitive to response distortion in Norwegian personnel selection situations. Consistent with our hypothesis, a large proportion, 31 of the 40 items, were answered differently among students under directed faking and normal response conditions, suggesting that high scores on the Good Impression scale might reflect an attempt to present self in a favourable way within a Norwegian context. In contrast to findings from other studies (Hough and Schneider 1996; McCrae and Costa 1983) and our hypothesis, the responses under faking conditions seemed to generalize to real-life applicant settings. The job applicants answered differently from the students under the normal test conditions on 31 of the items, and even obtained scores.
been suggested that the ability to self-enhance may be regarded as an aspect of social competence, and therefore analysis found that socially desirable responding was related to successful interactions at work. A meta-

high concern about the reactions of other people. It has appeared unrelated to performance.

When controlling for the portion of variance shared with external performance criteria, but rather was related to individual differences in conscientiousness and emotional stability (Ones, Viswesvaran and Reiss 1996). Consistent with our hypothesis, the variations in mean scale scores across the two cultures appeared to be unrelated to gender. Females in both cultures were as likely as their male counterparts to endorse desirable traits. Also, in examining the within culture gender differences in scale scores, we found no significant gender difference in either culture. These results suggest that the Good Impression scale is not likely to cause any adverse impact for either gender in personnel selection situations. However, more formal study is needed on the interaction between culture and gender in relation to self-presentation between the Norwegian and US subjects. In this study, the use of comparable student samples in Norway and the United States may have resulted in fewer and smaller cultural differences than might be observed with less educated or older adult samples.

The analyses of the Commonality scale suggests that mean differences in scores on the Good Impression Scale across the US and Norwegian samples did not occur as an artifact of acquiescence, which has been cited as one potential bias in cross-cultural comparisons. Analyses of response patterns on the two subsets of items in the Commonality scale that are conventionally answered in opposite directions, revealed no differences between Norwegian and US respondents either among men or women. For both subsets of items, more than 88% of respondents in both Norway and USA answered these items in the conventional direction. There were also no differences in the frequencies of respondents obtaining raw scores below 24, which is the established cut-off score for invalid protocols and indicative of random and erratic responses. Only one Norwegian respondent scored in this area of the scale, which is comparable to the rate found in the USA (Gough and Bradley 1996). Taken together, these results support our hypothesis that Norwegian students would not acquiesce more than their US counterparts and suggest that there are minimal differences in scale usage across the two cultures. It is worth noting, however, that a false–true format might be less sensitive to response style variance across cultures.
compared to scales involving larger response ranges (see Grimm and Church 1999).

In conclusion, results from this study indicate that the Good Impression scale can be employed reliably and validly in a Norwegian setting. Content analysis, interpretation and modification of the nine items that did not discriminate between normal and fake test instructions might represent a step to further improve the validity of the scale. The differences in mean scale scores between Norwegian and US students suggest that Norwegians may be less inclined to ascribe positive characteristics to themselves. Since we have no available data on the scores of US job applicants, whether or not these results generalize to a real-life employment setting requires further documentation. Differences in self-presentation might have important implications for the practicality of using personality measures in international selection settings, and such distinctions among Western societies could easily be overlooked. Because scores might carry different interpretations depending upon the cultural background of the candidates, direct comparison might be questionable and may result in unfair employment decisions. With expanding global markets, culturally diverse work teams, and expatriate work assignments, we urgently need more research on how to create selection systems that can cope effectively and fairly with candidates from different cultures.

Acknowledgements

This research was supported by grants from the Directory of Public Management in Norway. The authors are grateful for the support from ISU. They also want to thank Dr Dave Musson for valuable comments.

References


