Improving the Validity of Letters of Recommendation: An Investigation of Three Standardized Reference Forms

Julie M. McCarthy and Richard D. Goffin

Department of Psychology
The University of Western Ontario
London, Canada

Although letters of recommendation are widely used, little research has examined how accurately they predict job performance. The few existing studies have yielded mixed results, and meta-analytic estimates of validity range from .14 to .27 (Hunter & Hunter, 1984; Reilly & Chao, 1982). This investigation was designed to improve predictive validity by developing a standardized reference form and evaluating 3 different rating formats: Multi-Item scales, Relative Percentile Method (RPM) scales, and Global Trait Rankings. A total of 520 individuals applied to the Canadian military, and 544 LORs were obtained. Complete predictor and criterion data were available for 57 participants. Regression analyses indicated that the validity of the RPM rating format ($R_{(adj)}^2 = .18; R_{(adj)}^2 = .42$) was substantially higher than previous estimates of LOR validity. The 2 remaining methods produced nonsignificant results. Limitations of the study, suggestions for future research, and implications for the field are discussed.

The selection of suitable applicants is a critical component in producing and maintaining high levels of productivity in organizations. Although an impressive body of research has been conducted on most conventional selection methods, there is a noticeable lack of studies on letters of recommendation (LORs), despite the fact that they are commonly used (Aamodt, 1999). The paucity of empirical investigations suggests that there is much to learn about improving the predictive validity of LORs. Accordingly, our goal in this study was to develop a practical method of gathering applicant information through LORs that would maximize predictive validity.

Requests for reprints should be sent to Julie M. McCarthy, Department of Psychology, The University of Western Ontario, London, Ontario, Canada N6A 5C2.
Examination of existing literature reveals that there is no consensus on the term used to describe LORs. Reference check, referee report, reference statement, reference request, reference form, and letter of recommendation have all been treated as synonyms (Muchinsky, 1979). Moreover, the term reference is typically used to describe the individual who completes the LOR for an applicant. For the purposes of this article, an LOR refers to a description or evaluation of an applicant that is completed by an observer and used as a source of information for personnel selection. Although the term letter implies a narrative description of the applicant, observers may alternatively be required to present their views using a structured response format (Browning, 1968; Carroll & Nash, 1972; Daniel, 1990; Mosel & Goheen, 1958, 1959), as was the case in this study. In such cases the term letter is not, strictly speaking, correct; nonetheless, such research is routinely indexed using the marker letter of reference. The specific characteristics assessed in the LOR may vary, but typically include applicant knowledge, skills, abilities, and personality traits (Aamodt, 1999).

There are three primary reasons for including LORs as part of the overall selection process. The most frequently cited rationale is that they provide information on past performance, which has been found to be a good predictor of future success (Browning, 1968). Second, LORs may contribute unique variance to the overall selection process by acquiring information that is difficult to obtain from other sources (Aamodt, 1999). In particular, LORs can provide information about the applicant from the perspective of another person who has had the opportunity to observe the applicant’s personality and qualifications. The third notable advantage of LORs is that colleagues, peers, and coworkers serving as references have generally had the opportunity to observe the applicant’s typical and not just maximal performance. Research indicates that measures related to maximal performance (e.g., common selection tools such as resumes and unstructured interviews) may not provide an accurate representation of how well individuals will typically perform on the job (Sackett, Zedeck, & Folgi, 1988).

Although LORs have the potential to provide valuable applicant information, most suffer from leniency bias (Aamodt, 1999). The tendency to describe applicants in favorable terms is not surprising considering that the majority of applicants select their own references, and it is not difficult for a below-average candidate to find someone to write a positive LOR on his or her behalf (Aamodt, Bryan, & Whitcomb, 1993). Moreover, there are few rewards for providing an honest, but negative, evaluation of a candidate, whereas there may be serious legal penalties (e.g., defamation of character charges) if the applicant finds out (Aamodt, 1999). Leniency bias has been substantiated by Mosel and Goheen (1958), who found that almost 92% of ratings on occupational ability were either good or outstanding (on a 4-point scale: poor, satisfactory, good, and outstanding). Similar findings were reported by Browning (1968), who found mean reference ratings ranging from 3.22 to 3.48 on a 4-point scale. It is possible that these
high reference ratings are actually an indication that most of the applicants were superior and not evidence of leniency bias. However, a trend indicative of leniency has been observed in a number of studies (Aamodt et al., 1993; Carroll & Nash, 1972; Miller & Van Rybroek, 1988; Myers & Errett, 1959), leading Aamodt (1999) to suggest that such consistently high ratings can be viewed as evidence of leniency bias in LORs. Unfortunately, leniency can contribute to ceiling effects and a restricted range of LOR scores, which is likely to attenuate predictive validity estimates (Loher, Hazer, Tsai, Tilton, & James, 1997).

Notwithstanding the problem of leniency bias, LORs remain a ubiquitous selection tool (Aamodt, 1999). It is therefore surprising that our search uncovered only nine publications consisting of primary research aimed at assessing the predictive validity of LORs (Aamodt et al., 1993; Browning, 1968; Carroll & Nash, 1972; Clemente & Michener, 1976; Daniel, 1990; Mosel & Goheen, 1958, 1959; Stanton, Burstein, Kobos, & Loucks, 1979; Tremonti, 1973). In addition to these individual studies, two meta-analytic reviews (Hunter & Hunter, 1984; Reilly & Chao, 1982) have been conducted. Table 1 summarizes the main findings of these studies. Although the results of earlier investigations (Browning, 1968; Mosel & Goheen, 1958, 1959) suggest that LORs have low predictive validity, more recent studies (Aamodt et al., 1993; Carroll & Nash, 1972; Daniel, 1990; Hunter & Hunter, 1984; Stanton et al., 1979) suggest that moderate predictive validity is possible. An examination of Table 1 reveals that the highest validity estimates were obtained in the study by Aamodt et al. However, it is noteworthy that the procedure used by Aamodt et al. to score the LORs required an elaborate coding scheme, where two “experts” grouped all adjectives presented in the LORs into predefined categories. Although impressive predictive validity estimates were obtained, this scoring procedure required a labor-intensive, five-step process, making it impractical and less cost efficient for large-scale selection applications.

**THIS STUDY**

The purpose of this investigation was to improve the validity of LORs by developing a standardized reference form. A standardized form may be advantageous as it uses a structured format to which all references must adhere. This facilitates more consistent evaluation standards across applicants who have been rated by different reference writers. In addition, standardized formats may reduce the possibility that the characteristics of the reference writer (e.g., writing quality and style, personality) will spuriously influence the evaluation.

Three standardized formats were examined. The first was Multi-Item scales, which presented a series of personality-based items that were rated on a 7-point, Likert-type scale ranging from 1 (not at all descriptive) to 7 (extremely descriptive). The second was Relative Percentile Method (RPM) scales, which presented
# TABLE 1
Findings of Previous Criterion-Related Validity Studies

<table>
<thead>
<tr>
<th>Studies</th>
<th>Predictors</th>
<th>Sample</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mosel &amp; Goheen (1958)</td>
<td>Ratings from multiple LOR sources; five dimensions</td>
<td>1,117 employees in 12 skilled trades</td>
<td>.12—supervisory ratings</td>
</tr>
<tr>
<td>Mosel &amp; Goheen (1959)</td>
<td>Ratings from five LOR sources; multiple dimensions</td>
<td>400 employees</td>
<td>.13—supervisory ratings</td>
</tr>
<tr>
<td>Browning (1968)</td>
<td>Ratings from 11 LOR sources; six dimensions</td>
<td>508 teachers</td>
<td>.13—supervisory ratings</td>
</tr>
<tr>
<td>Carroll &amp; Nash (1972)</td>
<td>Forced-choice ratings of six items; 4-point Likert scale</td>
<td>98 clerical employees</td>
<td>.21—supervisory ratings; .18—turnover</td>
</tr>
<tr>
<td>Tremonti (1973)</td>
<td>Dean’s LOR rated by judges as excellent, good, or satisfactory</td>
<td>25 medical students applying for intern positions</td>
<td>Agreement was “no better than chance alone” (p. 277)</td>
</tr>
<tr>
<td>Clemente &amp; Michener (1976)</td>
<td>Dean’s LOR rated by judges for positive and negative content</td>
<td>107 medical students applying for intern positions</td>
<td>.01—medical staff ratings</td>
</tr>
<tr>
<td>Stanton, Burstein, Kobos, &amp; Loucks (1979)</td>
<td>Dean’s LOR rated by judges on a 5-point scale</td>
<td>12 medical students applying for intern positions</td>
<td>Significant correlation (p &lt; .05)* between dean’s LOR and peer ratings of performance</td>
</tr>
<tr>
<td>Reilly &amp; Chao (1982)</td>
<td>Several (meta-analysis)</td>
<td>5,718 applicants, multiple jobs</td>
<td>.14—several ratings</td>
</tr>
<tr>
<td>Hunter &amp; Hunter (1984)</td>
<td>Reference checks (meta-analysis)</td>
<td>Ranged from 415 (promotion) to 5,389 (supervisory ratings)</td>
<td>.26—supervisory ratings; .16—promotion; .23—training success; .27—tenure</td>
</tr>
<tr>
<td>Daniel (1990)</td>
<td>Ratings of a 15 adjective scale</td>
<td>39 work study students</td>
<td>.33—supervisory ratings</td>
</tr>
<tr>
<td>Aamodt, Bryan, &amp; Whitcomb (1993)</td>
<td>The number of traits appearing in a traditional LOR</td>
<td>104 graduate students</td>
<td>.32—overall grade; .38—teaching performance</td>
</tr>
</tbody>
</table>

*Note. LOR = letter of recommendation.

*The value of this correlation is not presented in this article.
a series of attributes that were rated on a 101-point scale ranging from 0 (below average) to 100 (above average). The third was Global Trait Rankings, which required references to rank order attributes from the most to the least characteristic of the applicant.

Each format was designed to combat the typical leniency bias associated with LORs. The Multi-Item scales were comprised of items that were written to be specific and to reflect observable behaviors. This rationale was borrowed from the performance appraisal domain, in which specific observable items have been used in an effort to increase rater accuracy (Murphy & Cleveland, 1995). The second format, RPM Reference scales, used percentile rating scales. It was anticipated that the use of a percentile scale might capitalize on a metric that has meaningful reference points and a wide range and thereby increase the variance in the references assessments of candidates. Although the RPM had not been previously used in the context of LORs, results using the RPM scale for performance appraisal have been encouraging (e.g., Goffin, Gellatly, Paunonen, Jackson, & Meyer, 1996; Jelley & Goffin, 2001; Wagner & Goffin, 1997). Finally, the forced-ranking procedure used in the Global Trait Rankings format circumvents range restriction because, much like a conventional forced-choice rating scale, the reference cannot give the candidate uniformly high scores on all dimensions. Previous research with a forced-choice reference format has proven to be successful (Carroll & Nash, 1972).

Our objective was to determine how well the three aforementioned formats performed as sets of predictors. Accordingly, the attributes included in each of the three formats were selected to reflect the entire predictor domain. This enabled our focus to be centred on the format and not the attribute level. To this end, empirical assessment of the three formats was conducted using forced-entry, multiple regression techniques, which retain all predictors in the final regression equation. Thus, the central question of this study was not, “What attributes predict military performance?” but rather, “Which format(s) in a standardized reference form can best predict performance in the Canadian Forces?”

In addition to testing the predictive validity of the previously mentioned formats, our research sought to avoid five major weaknesses present in previous LOR validity studies. First, in some past validational efforts, LORs were used to make hiring decisions. Under this methodology, candidates with high LOR ratings are more likely to be selected for the job. Similarly, those with low ratings are less likely receive a job offer. If accepted applicants possess predominantly high LOR scores, the range of LOR ratings is reduced, which may artificially reduce validity coefficients (Murphy & Davidshofer, 1998). Second, past studies have relied almost exclusively on simple bivariate correlations and have thereby not examined the overall predictive validity of multiple LOR dimensions. Third, job performance criteria used in past studies have typically been limited to performance appraisal measures that are used as the basis for actual administrative decisions (e.g., pay, promotions, and disciplinary action). This type of criteria is
likely to suffer from a variety of biases, many of which are political in nature (Murphy & Cleveland, 1995). For example, raters may deliberately inflate ratings to maximize a subordinate’s merit increases or to avoid the discomfort associated with giving negative feedback (Longenecker, Sims, & Gioia, 1987; Murphy & Cleveland, 1995). In contrast, raters may deliberately deflate ratings to shock a subordinate into a high level of performance or to send a message that the employee should consider leaving the organization (Longenecker et al., 1987). Fourth, many past studies failed to employ job analysis to ensure that the attributes to be assessed with the LOR are job relevant. Finally, both concurrent and predictive designs were employed in past research, but the latter design is more defensible. In predictive designs, actual job applicants are used as participants, and the data are collected in a personnel selection context. In contrast, concurrent designs utilize employees who have already been hired, and data collection does not occur in the context of actual selection. The use of concurrent designs is problematic for two reasons. First, employees may not be reflective of actual job applicants in that they may be less motivated to find letter writers who will be their most ardent supporters. This clearly limits the generalizability of findings. Second, existing employees have been through a personnel selection process and been identified as the most promising candidates. Accordingly, the range of LOR ratings may be reduced, which can artificially reduce validity coefficients (Murphy & Davidshofer, 1998).

We circumvented these difficulties in this study, as LORs were not used to make hiring decisions, multiple predictors were analyzed using regression, and the performance criterion was used for research purposes only. Further, job analysis was employed in the selection of the attributes to be measured, and actual job applicants served as participants within the context of a predictive validity design.

METHOD

Participants

The applicant pool consisted of 520 high school graduates from across Canada who applied to the 1997 Canadian Forces Basic Officer Training Program. Applicant ages ranged from 17 to 41 years old ($M = 20$ years old). The majority of applicants were men (73%). Of the 520 applicants, 359 were admitted into the program, resulting in a selection ratio of .69. Assuming two standardized reference forms per applicant, the total possible number of references was 1,040. In the case of 172 of the 359 accepted applicants, a single completed reference form was received, whereas in the case of 185 of the 359 accepted applicants, two completed reference forms were received. Overall, a total of 542 references completed the standardized reference forms.
References

Reference ages ranged from 17 to 75 years old ($M = 44$ years old), and the majority were men (69%). Most references (57%) were obtained from school settings (e.g., teacher, principal, academic counselor). An additional 21% of references were friends of the applicant. The remainder were from work settings (e.g., supervisors, coworkers), religious environments (e.g., clergy), or extracurricular activities (e.g., coaches). The average reference had between 6 and 9 years of experience writing reference letters ($SD = 1.63$), and 73.9% had successfully completed a university graduate or undergraduate program. The percentage of references reporting that they knew the applicant very well, well, and moderately well was 28.9%, 56.9%, and 14.3%, respectively. No references indicated that they knew the applicant only slightly or not well at all.

Procedure

From October 1996 until March 1997, Canadian Forces Recruiting Centers assessed the qualifications of applicants applying to the Basic Officer Training Program, a program providing military training and subsidized university education leading to officer positions in the Canadian Army, Navy, or Air Force. Applicant assessment for this program includes LORs, measures of cognitive ability and personality, structured interviews, biodata, and resumes. Standardized reference forms were obtained by instructing each applicant to provide the names and addresses of two references who were well acquainted with their knowledge, skills, and abilities. A package containing the standardized reference form and a return envelope was mailed to each reference. The standardized reference forms were not used in the selection process, and a statement to this effect was included on the LORs. Nonetheless, comments and feedback provided by the references indicated that they believed the LORs would be used for selection. Completed packages were returned to the researchers.1

In July 1997, selected candidates started the Basic Officer Training Program, an intensive 8-week program that provides training in leadership, field craft, weapons handling, navigation, and various other military skills. On completion of the course, peer performance ratings were collected for 238 of the candidates, each of whom had between one and seven peer raters. Supervisory ratings were obtained

---

1Several reference forms were not returned, hence applicants’ files contained one, two, or no completed forms. Where an applicant’s file contained only one form, the single reference ratings were used for all analyses. Where an applicant’s file contained two forms, the reference ratings were averaged to increase reliability.
approximately 8 weeks after the training course had ended. Supervisory performance ratings were ultimately obtained for 149 of the selected candidates.

This study took place in a bilingual environment where predictor and criterion data were made available in the respondents’ language of choice, French or English.²

Measures

**Standardized reference form.** The set of attributes to be assessed in the standardized reference form was selected on the basis of a job analysis conducted by two majors, eight military career counselors, one industrial or organizational psychology professor, and an industrial or organizational graduate student. The job analysis resulted in a list of 31 attributes, which were chosen to comprehensively encompass applicant characteristics that were relevant to performance in the target job and could reasonably be rated by a reference (see Table 2). As part of the job-analysis process, each attribute was rated in terms of its importance to the target job.

The standardized reference form contained three sections, each employing a different rating format. Sample items are presented in Figure 1. The first section contained six Multi-Item scales: Achievement, Emotional Stability, Intelligence, Leadership, Physical Fitness, and Team-Oriented. These scales were selected by job experts who took into account their importance to the target job (as derived through the job analysis) and the degree to which items for the scales could be observed by a typical reference. Each scale contained between 7 and 10 items. All items were behaviorally based and were rated on a 7-point Likert-type scale ranging from 1 (*extremely descriptive of the candidate*) to 7 (*not at all descriptive of the candidate*).

The second section of the standardized reference form contained RPM Reference scales, which were based on the work of Goffin, Gellatly, et al. (1996). This section required the reference to rate the applicant on the 31 attributes using an RPM rating scale. The RPM rating format requires that raters consider the performance of other candidates when evaluating the target individual. In this instance, references were instructed to consider each applicant relative to his or her peer group. Following standard RPM instructions (see Goffin, Gellatly, et al., 1996), ratings were made on a scale that ranged from 0 to 100, where 50 referred to the average performance of an applicant’s peer group. The RPM approach has been validated in two previous investigations (Goffin, Gellatly, et al., 1996; Wagner & Goffin, 1997), where it evidenced higher levels of accuracy and validity than individual performance appraisal methods. In addition, Jelley and Goffin (2001) found

²The observed results may underestimate the true relations, as this investigation was carried out in a bilingual environment where respondents had their choice of preferred language. It is conceivable that the observed relations would be stronger in a unilingual environment.
this format to be advantageous when used in conjunction with the Behavioral Observation Scale (BOS; Latham & Wexley, 1977) ratings. Specifically, they found that participants who made RPM ratings immediately before rating performance with the BOS scales were able to more accurately diagnose the target individual’s strengths and weaknesses with their BOS ratings.

The third and final section of the standardized reference form, Global Trait Rankings, required the rater to rank order 10 attributes from most descriptive to least descriptive of the applicant. Of the 10 traits, 6 were of particular substantive interest for predicting performance in the Basic Officer Training Program and corresponded to the six scales from the Multi-Item section. The remaining four attributes (Cooperativeness, Dominance, Emotional stability, and Energy level) were not of particular interest for this specific application.

<table>
<thead>
<tr>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic success</td>
</tr>
<tr>
<td>Accepting criticism</td>
</tr>
<tr>
<td>Achievement orientation</td>
</tr>
<tr>
<td>Attention to detail</td>
</tr>
<tr>
<td>Competitiveness</td>
</tr>
<tr>
<td>Control of anger</td>
</tr>
<tr>
<td>Cooperativeness</td>
</tr>
<tr>
<td>Decisiveness</td>
</tr>
<tr>
<td>Dominance</td>
</tr>
<tr>
<td>Emotional stability</td>
</tr>
<tr>
<td>Energy level</td>
</tr>
<tr>
<td>General responsiveness</td>
</tr>
<tr>
<td>Initiative</td>
</tr>
<tr>
<td>Innovation</td>
</tr>
<tr>
<td>Internal control</td>
</tr>
<tr>
<td>Leadership</td>
</tr>
<tr>
<td>Motivation toward Canadian Forces</td>
</tr>
<tr>
<td>Numerical ability</td>
</tr>
<tr>
<td>Oral communication</td>
</tr>
<tr>
<td>Orderliness</td>
</tr>
<tr>
<td>Perseverance</td>
</tr>
<tr>
<td>Physical endurance</td>
</tr>
<tr>
<td>Reasoning</td>
</tr>
<tr>
<td>Responsibility</td>
</tr>
<tr>
<td>Responsiveness to training</td>
</tr>
<tr>
<td>Self-esteem</td>
</tr>
<tr>
<td>Sociability</td>
</tr>
<tr>
<td>Team oriented</td>
</tr>
<tr>
<td>Traditional values</td>
</tr>
<tr>
<td>Verbal ability</td>
</tr>
<tr>
<td>Written communication</td>
</tr>
</tbody>
</table>
tiveness, Energy Level, Self-Esteem, Responsibility) were included as distractor variables that appeared to be of comparable desirability as the aforementioned six but were rated lower in job relevance. This rank-ordering procedure can be considered an extension of the forced-choice methodology as it required that raters discriminate among a group of dimensions that appeared equally desirable but had differential relevance to job performance (Carroll & Nash, 1972). Although the
Global Trait Ranking format resembles an ipsative scale (in which attribute scores are dependent on one another), it is not purely ipsative because four distractor variables were included. Purely ipsative scales suffer from multicollinearity, which attenuates the intercorrelations among scales and can bias predictive validity estimates (for a discussion of the psychometric limitations of purely ipsative scales, see Hicks, 1970). Accordingly, purely ipsative scales should be avoided when forced-choice formats are used in standardized reference forms.

The Multi-Item scales were presented first to reduce the possibility that raters would consider the more global RPM Reference scales, the Global Trait Rankings, or both, when completing the Multi-Item section. It is unlikely that this effect would be reversed—ratings from the Multi-Item scales influencing ratings on the other two sections—as items in the Multi-Item section were not organized according to the more global attributes. In addition, it is difficult for raters to base global ratings on randomly dispersed single items. The Global Trait Ranking section was presented last in an attempt to reduce premature rater attrition, as research indicates that raters dislike forced-choice rating formats (Cardy & Dobbins, 1994).

At the end of the standardized reference forms, references were given the opportunity to express their reactions to the LOR.

Performance criterion. An RPM Performance scale was developed and used for supervisor and peer ratings of officer candidate performance. Sample items are presented in Figure 2. There is a notable difference between this scale and the RPM-based reference ratings used in the standardized reference form. Specifically, the RPM Reference scales required the reference to rate one applicant on each percentile scale, whereas the RPM Performance scales required the rater to evaluate more than one candidate on each percentile scale.

The 12 scales comprising the RPM Performance scale were selected to provide comprehensive coverage of the job performance domain as determined by a Canadian Forces job analysis that was conducted by Bradley (1998). The scales include Technical Knowledge and Skill, Integrity, Physical Fitness, Communication, Cooperation, Leadership, Responsibility, Persistence, Quickness in Learning, Self-Esteem, Promotability, and Overall Performance. To avoid the biases associated with criteria that are linked to administrative decisions (as discussed earlier), these performance ratings were used for research purposes only.

RESULTS

Preliminary Analyses

With respect to the performance criterion measures, a total of 238 candidates had peer raters, 149 had supervisor raters, and 110 candidates had both. Ultimately, a
total of 57 candidates had usable data on the standardized reference form predictors and the performance criteria. There was no indication that loss of participants was nonrandom, as the demographic characteristics, predictor, and criterion scores of the reduced sample closely approximated those of individuals not included in the final analysis (see Table 3).

Consolidating the Predictor Domain

The analyses conducted in this section were based on the 357 applicants for which predictor data was available. The first step in consolidating the predictor domain was to ensure that the six Multi-Item scales were psychometrically sound. Therefore, item means, item standard deviations, internal consistency reliabilities, and
item–total correlations were examined. All items exhibited adequate statistical properties and were therefore retained. Internal consistency reliabilities ranged from .67 to .74 (n = 190–312). Scale intercorrelations ranged from low to moderate, with none approaching .71 (50% shared variance). This implies that references were able to distinguish the applicants’ performance on each of the scales.

The second section of the standardized reference form employed an RPM format and contained 31 attributes. To reduce the number of predictor variables in this section, a principal components analysis with varimax rotation was conducted (n = 306). On the basis of the scree plot, interpretability, and simple structure, a four-component solution was chosen (see footnote 3). The first component, labeled Military Adjustment, reflected the degree to which officers adapt to the military environment. Items loading on this component included Cooperation, Responsibility, Accepting Criticism, and Maintaining Composure. The second component, labeled Intelligence, reflected numerical, verbal, and reasoning abilities. Items loading on this component included Numerical Ability, Academic Success, Reasoning, Verbal Ability, and Achievement. The third component, labeled

---

3 A full-length technical report is available from the first author by request, which contains information on missing data, item means and standard deviations, scale internal consistency reliabilities, scale intercorrelations, RPM factor analysis, and correlations among predictor and criterion variables.
Physical Capabilities, reflected energy levels and motivation. Items loading on this component included Physical Endurance, Energy Level, and Competitiveness. The final component, labeled Leadership, reflected important leadership skills. Items included Dominance, Leadership, Oral Communication, Sociability, and Self-Esteem. Based on the loading matrix, items were unit weighted and aggregated to form the four RPM Reference scales.

The third section of the standardized reference form, Global Trait Rankings, required the rater to rank order 10 attributes from most descriptive to least descriptive of the applicant. As previously mentioned, six traits were of particular substantive interest for predicting performance in the target job. Descriptive statistics on these rankings (n = 350) revealed low standard deviations (2.41–2.72) and low interscale correlations (.02–.36). The low interscale correlations make intuitive sense, as the rank-ordering procedure does not allow consistently high or low ratings across all six dimensions for any given candidate, which ultimately results in lower intercorrelations among the scales.

Consolidating the Criterion Domain

The analyses conducted in this section were based on 238 applicants for whom peer performance ratings were available and 149 applicants for whom supervisory performance ratings were available. As previously indicated, peers and supervisors rated the performance of each officer candidate on the 12 RPM Performance scales. To reduce the number of variables, principal components analyses with varimax rotation were conducted on both the peer and supervisory performance ratings. Results were unequivocal; the scree plots and substantive interpretation strongly suggested a single-factor solution for both the peer and supervisory performance scales. Accordingly, the 12 RPM scales were unit weighed and aggregated to create “total peer” and “total supervisor” performance measures. The correlation between the total Peer and Supervisory rating scales was positive and substantial (r = .51, p < .01, n = 108), suggesting that it would be beneficial to aggregate these scales (Paunonen, 1984). Thus, the overall peer and supervisory ratings were unit weighted and combined to create a single performance criterion.

Predictive Validity of the Standardized Reference Form

Correlations. Correlations between the standardized reference form predictors and the performance criteria are presented in Table 4. As shown, the relations involving the Multi-Item scales ranged from –.07 to .25 and were nonsignificant. In the case of the RPM Reference scales, the Military Adjustment dimension was sig-
significantly correlated with the criterion \( r = .30, p < .05 \). The other three relations involving the RPM Reference scales were nonsignificant but were in the positive direction. The Global Trait Rankings gave rise to a significant, but negative, relation between the Intelligence dimension and the RPM Performance criterion \( r = –.33, p < .01 \). This implies that applicants for whom Intelligence was ranked higher than the other five dimensions tended to receive lower performance scores. The five remaining relations between the Global Trait Rankings and the performance ratings were nonsignificant.

**Regression analyses.** A series of regressions was carried out to determine how well each of the three formats of the standardized reference form predicted the performance criterion and whether the three formats demonstrated incremental validity over each other in the prediction of the criterion. An assessment of incremen-
tal validity is particularly important, as it provides insight into what combination of formats or what single format is most advantageous for selection purposes. Although the predictors to be included in the multiple regression could have been chosen based on zero-order correlations, as previously mentioned, our goal in this study was to assess how well the three sections performed as complete sets of predictors. In addition, selecting predictors from a larger set on the basis of their correlations with a criterion artificially inflates $R^2$ and produces shrinkage on cross-validation (Cohen & Cohen, 1975). Therefore, a forced-entry approach was used to ensure that the specified predictors were retained in the regression equation, irrespective of the significance level of their beta weights. In other words, all variables for a measure were treated as a set, and the various sets were added to the regression analysis in a specified order.

The regression results are presented in Table 5. To provide a complete picture of the results, both unadjusted and adjusted (adj) $R^2$ values are presented; however, emphasis is placed on the adjusted $R^2$ estimates as they provide a better estimation of the respective population value. In Step 1, single-step regressions were carried out to determine how well each of the three formats predicted the performance criterion. Of these, only one was significant, the RPM Reference scales ($R^2 = .24, p < .01; R^2_{(adj)} = .18$).

In Step 2, sets of predictors, each corresponding to one of the three formats, were entered in combination with other sets to assess incremental predictive valid-

<table>
<thead>
<tr>
<th>Regression</th>
<th>$R^2$</th>
<th>$R^2_{(adj)}$</th>
<th>$R^2$</th>
<th>$R^2_{(adj)}$</th>
<th>$R^2$</th>
<th>$R^2_{(adj)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Multi-Item scales first, RPM scales second</td>
<td>.11</td>
<td>.00</td>
<td>.26</td>
<td>.10</td>
<td>.15</td>
<td>.10</td>
</tr>
<tr>
<td>2. RPM scales first, Multi-Item scales second</td>
<td>.24*</td>
<td>.18</td>
<td>.26</td>
<td>.10</td>
<td>.02</td>
<td>.08</td>
</tr>
<tr>
<td>3. Multi-Item scales first, Global Trait Ranking scales second</td>
<td>.11</td>
<td>.00</td>
<td>.21</td>
<td>.00a</td>
<td>.10</td>
<td>.00a</td>
</tr>
<tr>
<td>4. Global Trait Ranking scales first, Multi-Item scales second</td>
<td>.13</td>
<td>.03</td>
<td>.21</td>
<td>.00a</td>
<td>.08</td>
<td>.00a</td>
</tr>
<tr>
<td>5. RPM scales first, Global Trait Ranking scales second</td>
<td>.24*</td>
<td>.18</td>
<td>.39*</td>
<td>.26</td>
<td>.15</td>
<td>.08</td>
</tr>
<tr>
<td>6. Global Trait Ranking scales first, RPM scales second</td>
<td>.13</td>
<td>.03</td>
<td>.39*</td>
<td>.26</td>
<td>.26*</td>
<td>.23</td>
</tr>
</tbody>
</table>

Note. Sample sizes ranged from 57 to 60. adj = adjusted; RPM = Relative Percentile Method. *Actual application of the multiple $R$ formula resulted in a negative value, which is not possible in the population; thus 0 was substituted. *$p < .01$.  

**TABLE 5**
Regression Analyses

Squared Multiple Correlations

<table>
<thead>
<tr>
<th>Regression</th>
<th>$R^2$</th>
<th>$R^2_{(adj)}$</th>
<th>$R^2$</th>
<th>$R^2_{(adj)}$</th>
<th>$R^2$</th>
<th>$R^2_{(adj)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Multi-Item scales first, RPM scales second</td>
<td>.11</td>
<td>.00</td>
<td>.26</td>
<td>.10</td>
<td>.15</td>
<td>.10</td>
</tr>
<tr>
<td>2. RPM scales first, Multi-Item scales second</td>
<td>.24*</td>
<td>.18</td>
<td>.26</td>
<td>.10</td>
<td>.02</td>
<td>.08</td>
</tr>
<tr>
<td>3. Multi-Item scales first, Global Trait Ranking scales second</td>
<td>.11</td>
<td>.00</td>
<td>.21</td>
<td>.00a</td>
<td>.10</td>
<td>.00a</td>
</tr>
<tr>
<td>4. Global Trait Ranking scales first, Multi-Item scales second</td>
<td>.13</td>
<td>.03</td>
<td>.21</td>
<td>.00a</td>
<td>.08</td>
<td>.00a</td>
</tr>
<tr>
<td>5. RPM scales first, Global Trait Ranking scales second</td>
<td>.24*</td>
<td>.18</td>
<td>.39*</td>
<td>.26</td>
<td>.15</td>
<td>.08</td>
</tr>
<tr>
<td>6. Global Trait Ranking scales first, RPM scales second</td>
<td>.13</td>
<td>.03</td>
<td>.39*</td>
<td>.26</td>
<td>.26*</td>
<td>.23</td>
</tr>
</tbody>
</table>
ity. This produced six hierarchical regressions (see Table 5). Only the combined set of RPM and Global Trait Ranking scales, the fifth and sixth regressions, resulted in a significant $R^2 (R^2 = .39, p < .01; R^2_{(adj)} = .26)$. In terms of incremental validity (the final column of Table 5), a significant increase in variance occurred in Regression 6, where the RPM Reference scales were entered following the Global Trait Rankings ($\Delta R^2 = .26, p < .01; \Delta R^2_{(adj)} = .23$). The change in $R^2$ was not significant when this order was reversed.

**DISCUSSION**

In this investigation we examined whether a standardized reference form could predict performance in the Canadian Forces and, more specifically, which of three standardized rating formats provided optimal prediction. The results indicate that RPM Reference scales can significantly predict military performance ($R^2_{(adj)} = .18$). This adjusted multiple correlation translates into a validity coefficient of approximately .42, which compares favorably to the reported validities of many conventional selection methods (Schmidt & Hunter, 1998). Further support for the RPM format was found in the references’ comments, with more than 60% of references indicating that they were very confident in completing the RPM Reference scales.

Findings also indicate that a leniency effect might be occurring for the RPM Reference scales, as the average scores were high (80 on a 0–100 scale). As discussed, leniency bias produces range restriction in reference ratings and can attenuate criterion-related validity. To this end, the crucial consideration becomes the extent to which ceiling effects occurred. For the RPM Reference dimensions, references rated applicants’ performance at the top of the percentile scale only 4.3% of the time. This implies that ceiling effects were not likely to be pervasive enough to attenuate the correlations and further supports the viability of the RPM Reference scales.

In contrast to the RPM Reference scales, the Multi-Item scales were unsuccessful in predicting performance in the Basic Officer Training Program ($R^2_{(adj)} = .00$). Results did indicate, however, that the Multi-Item scales shared a substantial amount of variance with the RPM Reference scales, which is evidenced by the finding that the RPM Reference scales did not add significant incremental validity to the Multi-Item scales. The most likely explanation for the shared variance is that the same underlying dimensions were reflected in the RPM and Multi-Item scales.

One possible explanation for the lack of predictive validity in the Multi-Item scales is that a restriction of range may have occurred. The Multi-Item scales means were extremely high (5.9 on a 7-point scale). Moreover, references rated applicants’ performance at the top of the 7-point scale 34.3% of the time. This is considerably higher than the respective 4.3% estimate found in the RPM Reference scales, and therefore, there is a greater likelihood that a ceiling effect oc-
curred for the Multi-Item scales. This may help to explain the higher predictive validity found for the RPM Reference scales.

The third and final section, Global Trait Rankings, also showed little promise. Regression analyses indicated that the Global Trait Ranking format failed to predict the candidates’ performance ($R^2_{(adj)} = .03$). Results also revealed that a significant prediction was found when this section was combined with the RPM Reference scales ($R^2_{(adj)} = .26$); however, the Global Trait Rankings did not add incremental validity beyond that of the RPM Reference scales, although the converse was true. This implies that using a combination of the RPM and Global Trait Ranking formats will not improve the predictive validity beyond that already obtained using the RPM Reference scales on their own. The failings of the ranking format were further evidenced by the profusion of negative comments provided by the references. Only 37.1% of references were “very confident” in completing this section, and many provided written comments that clearly expressed their dislike for the rank-ordering procedure.

Overall, the results of this study suggest that the traditional view that LORs are of little practical value (Reilly & Chao, 1982) may be premature. Although meta-analytic results have found validity estimates ranging from .14 to .27 (Hunter & Hunter, 1984; Reilly & Chao, 1982), this study demonstrated a validity coefficient of .42 for the RPM Reference scales. This difference is substantial, particularly when considered in light of the fact that the RPM format is more practical and cost efficient than more labor-intensive methods that have yielded slightly lower estimates (e.g., Aamodt et al., 1993; Peres & Garcia, 1962). Nonetheless, we expect that Aamodt et al.’s approach is likely to remain an attractive option in cases in which narrative LORs continue to be used.

**Strengths and Limitations**

As indicated in the introduction, in designing this investigation we used several techniques to overcome the methodological shortcomings of previous validity studies as well as the common applied problems with using LORs. The main strength of this work is the assessment of actual job applicants within the context of a predictive validity design. Additional strengths include use of a job analysis to select job-relevant attributes, use of reference forms for research purposes only to avoid range restriction in applicant scores, collection of performance ratings for research purposes only to avoid some of the biases associated with operational performance appraisal assessments, inclusion of written items that reflected observable behaviors to reduce the potential for leniency bias, and use of rating formats that could reduce the possibility of leniency bias (a percentile scale ranging from 0 to 100 and a forced-ranking scale).
Despite the aforementioned strengths, it is important to note that this study contains three weaknesses. The first centers on the sample size for the predictive validity analyses. Most researchers are all too familiar with the challenge of securing large samples for predictive validity designs (Schmidt & Hunter, 1981). Unfortunately, this difficulty is clearly illustrated in this investigation, which began with a sample of 520 applicants and resulted in only 57 participants for the multiple regression analyses. This number exceeded the minimum sample size requirement for multiple regression, loss of participants appeared to be random, and a power analysis of the multiple regression analysis indicated that the power was greater than .80 (assuming $k = 4–6$ and $R^2 = .24$). Nevertheless, this is a considerable reduction from the original 520 applicants, for which there are three main reasons. First, 151 participants were lost due to hiring decisions (not based on the standardized reference form). Second, a nonoperational criterion measure was employed, making it difficult to obtain high response rates, as it was not part of formal organizational procedures. Third, the criterion analysis for this investigation required that data from three separate sources were available for each target: reference ratings on the standardized reference form, supervisory performance ratings, and peer performance ratings. Future researchers attempting to carry out multisource research such as this would be well advised to allow for considerable loss of participants.

A second limitation is that there may be common method variance for the RPM Reference scales predicting the RPM Performance scales. Although in this study it was not possible to empirically examine whether this occurred, there are two reasons why this potential effect is considered minimal. First, the two measures were completed by two different sets of people—the RPM Reference scales were completed by the references, and the RPM Performance scales were completed by the candidates’ supervisors and peers. Second, although the basic RPM format was the same, there was a notable difference between the task required of the reference and that required of the supervisor and peer raters. Specifically, the RPM Reference scales required each reference to rate only one applicant and were therefore likely to have fostered serial information processing. In contrast, raters completing the RPM Performance scales were required to rate more than one applicant at a time, which was likely to foster parallel processing. Therefore, although shared method variance is a possibility, it seems unlikely that the predictor and criterion measures shared a significant amount of systematic method variance.

The third limitation of this investigation is that a method–construct confound may have influenced the results. Specifically, the goal of this study was to evaluate three different LOR formats; however, the content of the three formats was not identical. Although the Multi-Item scales and Global Trait Rankings contained the same six dimensions, the RPM Reference scales included 31 attributes, which were later consolidated into four dimensions. The nature of the formats rendered it infeasible to
include all 31 attributes identified in the job analysis in each section of the standardized reference form. Specifically, the methods were so different by nature that if the same constructs were measured for all three methods, a confound would have existed in terms of the amount of time that it took the references to fill out all three formats. In other words, references would have spent a disproportionate amount of time completing Multi-Item and Global Trait Ranking scales. From an applied perspective, standardized administration time is an important consideration, because if the Multi-Item or Ranking approaches were actually used for selection purposes, they would be limited to measuring fewer constructs than would the RPM approach. Accordingly, a fair test of all three approaches could only be carried out if the time required to complete the three methods was roughly equivalent.

Therefore, to obtain consistency in administration time, we included only six Multi-Item scales. In contrast, the RPM Reference scales are global ratings and require a period of time that is analogous to that needed to fill out a single item, so it was therefore possible to assess all 31 attributes in this section. Consolidation of these attributes resulted in a four-dimensional structure, three of which (Intelligence, Leadership, and Physical Capabilities) directly correspond to three of the Multi-Item scale and Global Trait Ranking dimensions (Intelligence, Leadership, and Physical Fitness). The fourth RPM dimension, Military Adjustment, reflected control of anger, cooperation, composure, and motivation toward the Canadian Forces and is similar to the Multi-Item and Global Trait ranking Emotional Stability dimension. The final format, Global Trait Rankings, required references to rank order a list of attributes. Clearly, asking applicants to rank 31 attributes would have been an unwieldy task. Accordingly, the same six attributes that were used in the Multi-Item scales were assessed with this format. Moreover, the subordinate dimensions underlying the 31 RPM attributes resembled four of these same six attributes. Overall, this strategy removes the possible confound due to the amount of time spent on the three different scales and directly reflects what would occur if these approaches were used in applied settings. Therefore, although the possibility of a method–construct confound cannot be discounted, we consider this approach to be the most appropriate test of the three LOR formats.

Suggestions for Future Research

Focusing largely on prediction, in this study we were unable to address the construct validity of the predictors thus employed. Future research aimed at establishing the nomological network of RPM predictors would be an important first step in this direction (see Murphy & Davidshofer, 1998), as would research on whether these results are limited to the specifics of our selection scenario. Thus, cross-validation using other target jobs and other organizations is needed. Also of interest is the extent to which the standardized reference approach promoted here provides in-
incremental validity beyond that of other common selection tools such as traditional LORs, interviews, and personality testing.

An additional avenue for future investigations is the cognitive processes that LOR readers engage in when interpreting these letters. Such an investigation might provide valuable insight into the enhancement of letter readers’ accuracy in making selection decisions and suggest potential avenues for the training of letter readers. A recent article by Loher et al. (1997) introduced a communication-based theory for LORs and provides an excellent framework for future research in this area.

Implications for the Field

The use of valid selection procedures can result in significant increases in organizational productivity, commitment, and satisfaction, as well as decreases in turnover, absenteeism, and training expenses (Cascio & Thacker, 1994)—thus reinforcing the importance of developing techniques to improve the predictive validity of one of the most pervasive selection tools, the LOR. In this study we demonstrated that the use of a standardized reference form employing an RPM format significantly predicted officer performance in a Basic Officer Training Program. If future research substantiates these findings, then organizations may wish to consider the use of standardized reference forms with RPM formats as part of their selection process.

The results of this study also provide further support for the RPM rating format. Although this format has been used for performance assessment in several past publications (i.e., Christiansen, Goffin, Johnston, & Rothstein, 1994; Gellatly, Paunonen, Meyer, Jackson, & Goffin, 1991; Goffin & Gellatly, 2001; Goffin, Gellatly, et al., 1996; Goffin, Rothstein, & Johnston, 1996, 2000; Jelley & Goffin, 2001; Meyer, Paunonen, Gellatly, Goffin, & Jackson, 1989; Wagner & Goffin, 1997), this investigation was the first to incorporate it into a standardized reference form. Not only were the RPM Reference scales able to significantly predict performance, but the RPM format was well received by raters and proved to be less labor intensive than other approaches.

Summary

This study is the first to investigate the predictive validity of three different rating formats for use in a standardized reference form. We used several techniques to overcome the methodological shortcomings of past research as well as the common applied problems in using LORs. The results of this study indicate that the RPM format demonstrated substantial predictive validity when used in a standardized reference form. It is hoped that the results of this study will serve as a catalyst for future research on standardized reference forms.
ACKNOWLEDGMENTS

Partial funding for this project was provided by three sources: the Canadian Military, a doctoral fellowship from the Social Sciences Research and Humanities Council of Canada held by Julie M. McCarthy, and Grant 410–990–0247 from the Social Sciences Research and Humanities Council of Canada held by Richard D. Goffin. A preliminary version of this article was presented at the 14th Annual Meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA, April 1999. There are several individuals in the Canadian Military who helped make this project possible, including Peter Bradley, Lisa Noonan, and Jim Yuchiyama. To them we extend our sincere thanks. We also gratefully acknowledge the editor and two anonymous reviewers for their many insightful suggestions.

Although support for this work is gratefully acknowledged, we would like to note that the opinions expressed in this article do not necessarily reflect those of the Canadian Military.

REFERENCES

istration, 35, 42–46.
tional interests, and cognitive predictors of managerial job performance and satisfaction. Personality and Individual Differences, 12, 221–231.


