

Employment References: Who Are We Talking About?

Michael G. Aamodt, Mark S. Nagy, & Naceema Thompson

Radford University

Department of Psychology

Radford, VA 24142

(540) 831-5513

maamodt@runet.edu

Reference checks and letters of recommendation are commonly used techniques in employee selection. References are typically obtained to confirm details on an application, check for prior discipline problems, discover new information about an applicant, and predict future performance.

Though employment references are a commonly used personnel selection technique, research has not been supportive of their validity. The purpose of this paper is to explore the possibility that characteristics of the reference provider contribute to this lack of validity.

Problems With References

Leniency

One of the biggest problems that plague references is that they tend to be inflated, or at least excessively positive (Reilly & Chao, 1982; Miller & Van Rybroek, 1988; Range, Menyhert, Walsh, Hardin, Ellis & Craddick, 1991). The reasons for this problem are many. First and foremost, applicants get to choose who writes their letters; thus, recommendation letters probably reflect the *best* qualities about an applicant rather than the *average* or *overall* qualities of an applicant. Second, references and recommendation letters may be inflated because reference providers may fear legal retribution or, at the very least, a heated confrontation with the applicant should the reference provider state something that may be interpreted as a negative feature (Aamodt, 1996). Evidence of this concern is provided when one discovers that letters written for applicants who did *not* waive their right to see the letter (that is, applicants did have access to the letter) were rated more positive than letters written for applicants who *did* waive their right to see the letter (Ceci & Peters, 1984).

As a result of these problems, several researchers have recommended that reference providers include both positive and negative information about the applicant. However, as noted by Range et al., (1991), these suggestions will be ineffective until they are incorporated universally; given the widespread use of references and recommendation letters, this scenario is highly unlikely.

Writer Influences

Another troubling aspect of recommendation letters is that several of the problems linked to their use may be the fault of the actual writer of the letter. For example, Ralston and Thameling (1988) reported that vividness of language (which was essentially the use of specific examples) was positively related to the perceived favorableness of the information contained in recommendation

letters. Similarly, Knouse (1983) found that specific examples enhanced positive perceptions of the applicant as well as the credibility of the letter writer. In a similar vein, previous research has indicated that writers compose longer letters for persons they like more (Mehrabian, 1965) and that longer letters are perceived to be more favorable (Kleinke, 1978). Given that writers may naturally differ in their approach to writing letters (i.e., some may be more specific and/or write longer letters than others), it follows that the success of an applicant's letter may be due, in part, to what type of writer has written the recommendation letter.

This notion that the value of a letter may depend on who writes it is strengthened when one considers the low reliability typically reported for letters of recommendation. Typically, reported reliabilities of letters of recommendation are under .40 (Aamodt, Dwight, & Michals, 1994). In other words, two different letters written about the same person are not very consistent in their appraisal of performance. In fact, previous research indicates that the content of the letters may reveal more about the letter writer than the actual applicant. In a study that examined letters of recommendation to graduate school, Baxter, Brock, Hill, and Rozelle (1981) found that letters written by the same person about two different applicants were actually more consistent in content than letters written by different persons for the same applicant.

When considered with the influences of specific examples and length of letters, the finding that letters may be more consistent *within* raters (rather than between raters) certainly suggests that letters of recommendation may indeed say more about the author than the applicant. Although this finding may have important implications for the use of recommendation letters, there have been no replications of the Baxter et al. (1981) study to date. In addition, Baxter et al. used a coding strategy that involved a list of "psychological or inferential qualities" (Baxter et al., p. 297), of which they provided only a few examples. However, when investigating letters of recommendation, a more fruitful approach may be to rely on the coding system of Peres and Garcia (1962), which lists various adjectives explicitly developed for scoring letters of recommendation.

The Present Studies

To investigate the role of the recommender, two studies were conducted. The first study was designed to test the idea that information in letters of recommendation says more about the person writing the letter than about the person being written about. The second study tested the idea that a less favorable perception would be given to applicants whose reference providers committed grammatical errors.

Hypothesis 1: Consistent with previous research, there will be little agreement between two writers of letters for the same applicant.

Hypothesis 2: Consistent with Baxter *et al.* (1981), there will be significant agreement between the traits by used one writer to describe two separate applicants.

Hypothesis 3: When many grammatical errors are committed by a referee, less favorable perceptions of the applicant will occur than when few or no grammatical errors are committed by the referee.

Study 1

Method

Participants. Two samples were used in the study. The first sample consisted of 155 graduate students who had received master's degrees in psychology. These students were part of a much larger data set used to validate a psychology department's admissions criteria. At least two letters of recommendation were available for each of the 155 students. The second sample consisted of 127 university faculty who, in the past five years, had written letters of recommendation for at least two students applying to the department's master's degree program in psychology.

Procedure. The letters of recommendation were analyzed using the trait method created by Peres and Garcia (1962) and refined by Aamodt, Bryan, and Whitcomb (1993). For each letter of recommendation we

1. Highlighted the traits for each letter.
2. Used the trait list found in Aamodt (1996) to place each of the traits into one of five categories. The original Peres and Garcia categories were mental agility (e.g. smart, analytical, one of my best students), vigor (e.g. motivated, dynamic, energetic), urbanity (e.g. outgoing, extroverted, good social skills), cooperation-consideration (e.g. helpful, agreeable, loyal), and dependability-reliability (responsible, conscientious, always attended class). In keeping with current thinking on personality, we used the "Big 5" dimensions of openness, extroversion, agreeableness, and conscientiousness, and kept the Peres and Garcia vigor dimension. Although the vigor dimension does not have a Big 5 counterpart, it is consistent with Hogan Personality Inventory division of extroversion into two categories. Further, the Big 5 dimension of emotional stability is rarely mentioned in letters. Previous research had indicated that the interrater reliability for placing traits into the five categories is .92 (Aamodt, Dwight, & Michals, 1994).
3. To control for effects of letter length and number of traits used, we divided the number of traits in each category by the total number of traits across the five categories.
4. Counted the total number of lines in the letter of recommendation starting after the salutation (e.g., Dear Admissions Committee) and ending prior to the signatory closing (e.g., cordially, sincerely).

Once the letters in Sample 1 were scored, the trait percentages in each of the five dimensions for one referee were correlated with the trait percentages in each of the five dimensions for the second referee. For the Sample 2 letters, the trait percentages in a letter written by a referee for one student were correlated with the trait percentages in a letter written by the same referee for a different student. To determine the significance of the differences between the correlations in each trait dimension, Fisher's Z_r transformation was used.

Results and Discussion

As shown in Table 1, our hypotheses were generally supported as the traits used by a referee to describe one applicant were significantly correlated with the traits used by the same referee to describe another applicant for three of the five trait categories (openness, $r = .44$; vigor, $r = .19$; and conscientiousness, $r = .32$). The traits used by two referees to describe the same applicant were significantly correlated for only one of the five trait categories (conscientiousness, $r = .31$, $p < .01$). Thus, there was greater consistency in the traits used by one person to describe two applicants than in the traits used by two people to describe one applicant. These results appear to support the idea that letters of recommendation say more about the person writing the letter than they do about the person being written about (Baxter et al., 1981).

Table 1

Correlations between traits used by one letter writer to describe two applicants and the traits used by two letter writers to describe the same applicant.

Trait Category	One writer describing two applicants (N = 127)	Two writers describing one applicant (N = 155)	Z
Openness	.44**	.05	3.49**
Vigor	.19*	.11	.68
Extroversion	.15	.07	.67
Agreeableness	.10	.11	.08
Conscientiousness	.31**	.20*	.98
Total lines	.64**	.16*	4.93**

As also shown in Table 1, the correlation between the length of letters written by one person for two applicants were highly correlated ($r = .64$, $p < .01$). This is an interesting finding as previous research has indicated that people reading letters of recommendation perceive that longer letters indicate that the referee liked the applicant (Mehrabian, 1965). Our findings suggest that the length of a letter of recommendation is actually more a function of the letter writer's style than it is an indication of attitude toward the applicant. In addition, our results are different from those of Mehrabian who found that subjects, told to write a positive letter of recommendation for a person they liked, wrote longer letters than subjects writing positive letters for people they didn't like. The difference in finding might be due to research design differences. For instance, Mehrabian asked undergraduates to write letters of recommendation whereas our study examined actual letters of recommendation written by professionals for actual applicants. Moreover, Mehrabian's study had the students imagine a situation in which they would be writing a letter for someone they either liked or disliked; our study used no such imagination manipulation. Finally, Mehrabian instructed students to write positive letters; our study had placed no such limitation on our letter writers (indeed, some of the letters were negative). In sum, our study not only used a more applied sample, but our results indicate that length of the letter may have more to do with the actual writer's style than

the degree of attraction between the letter writer and the applicant.

Our finding that the content of recommendation letters may have more to do with the person writing the letter than with the applicant has several implications for using letters of recommendation in selection. First, if letters are used to select applicants then one must consider not only the applicant, but the source of the letter. Traditionally, the selection process is designed to collect information about the applicant only. However, our research indicates that using letters of recommendation collects information about the applicant *as well as* information about the letter writer. Thus, when considering recommendation information, the evaluator should consider the source of the information.

Of course, it would be extremely difficult to consider the source of the letter writer unless some sort of comparison could be used. Thus, a second implication of our findings would be to suggest that evaluators not only keep letters of recommendation about applicants on file (a common practice), but also keep information on *who wrote the letters*. If this information were on file, then evaluators could refer to previous letters written by the same author and then make a comparison between the two letters. If one of the letters was judged to be more positive, then this information could be used in the selection decision. Furthermore, we suggest that the more two applicants have in common, the more helpful this comparison would be. For example, judging letters written by the same person for two different applicants applying to graduate school in *different* years could be more helpful than if one did not consider the letter writer at all. Moreover, judging letters written by the same person for two different applicants applying to graduate school in the *same year* would be even more helpful. Thus, it appears that considering who wrote the letter could be an important piece of information, particularly when the applicants are have something in common (such as applying for the same position and/or applying at the same time).

Finally, an argument could be made that a strength of recommendation letters is that each letter provides different information about the applicant. For instance, a popular reason given for the low reliability of recommendation letters is that the different writers do not observe the same applicant in the same context (Aamodt, 1996). That is, one letter writer only observes one aspect of the applicant (e.g., in a classroom) whereas other letter writers may observe other aspects of the applicant (e.g., as a research assistant or an advisee). Although this differential observation yields low reliability (and consequently is thought to be an indication of the low usefulness of recommendation letters), it may be that different letter writers are be giving a more "complete picture" of the applicant by providing different information. This is essentially thought to be one of the strengths of a relatively new trend in performance appraisal often referred to as "360 degree feedback". In this process, it is assumed that a more complete picture of an employee's performance is gained when using multiple sources (such as peers and customers) than only one source (i.e., an employee's supervisor) to evaluate performance. In fact, some research has indicated that these different sources do indeed provide different information regarding performance (Erismann & Reilly, 1996). In much the same way, it could be argued that letters of recommendation provide different sources of information about an applicant, and that this information is *helpful* in the selection process, despite the low interrater reliability it often yields.

Finally, considering letters of recommendation in the same context as a "360 degree" format only enhances our findings that the same letter writer is more consistent across applicants when it is realized that the same writer often serves in the same role for many applicants (e.g., a professor). In other words, if a supervisor was able to observe her employee in the context of many roles (e.g., as a peer or as a customer), then the supervisor alone could probably provide a more complete picture of an employee. In much the same way, if a letter writer could observe the applicant in a variety of contexts (e.g., as a professor and as a friend and as a supervisor), then the letter writers could probably provide a more complete picture of the applicant and consequently the letters written would be more consistent *across* writers. However, this is obviously not the case, and thus the letter writer usually can only observe many applicants in the same role. As a result, the applicants may be described more consistently *within* writers because they tend to serve in the same role to the applicant.

The results of study one indicate that there is greater agreement between two letters written by one person for two applicants than there is between letters written by two people for the same applicant. These findings suggest that a potential reason for the low reliability of recommendation letters is that writers write similar letters for all applicants. Further research is needed to determine if this similarity is a function of the writer's own personality, the writer's "schema" of what a recommendation letter should be like, or some other explanation.

Study 2

Method

Participants. Participants were 217 students (82% freshmen) enrolled in six sections of introductory psychology. Each class was randomly assigned to one of the six experimental conditions. Fifty-eight percent of the participants were female. Ages ranged from 18 to 30 with a mean of 18.

Procedure. Subjects were randomly assigned to one of six experimental conditions formed by a three (grammatical errors: 0, 2, 6) by two (job status: architect, construction worker) factorial design. Half of the subjects were told that they would hear a tape of a reference check being conducted on an applicant for an architect position (high status) and the other half told they would hear a tape of a reference check conducted on an applicant for a construction worker position.

Each tape contained the same information but varied in the number of speech errors committed by the reference provider; zero, two, or six. Speech errors were defined as any mistake made using the English language. Types of errors included mistakes in grammar, using verb tenses incorrectly, and using words inappropriately.

Participants rated their perception of the applicant on a three-item questionnaire. The three questions employed a Likert format with five-point scales and anchors at each end of the scale (1 = poor, 5 = excellent). A fourth question was used as a manipulation check to see if participants attended to the presence or absence of the speech errors. Only the 153 (70%) participants who

passed the manipulation check were included in the analysis.

Results

The results of a MANOVA supported Hypothesis 3, which stated that applicants whose reference providers committed six speech errors would receive lower favorability ratings than those applicants whose reference providers committed no speech errors, $F(6, 290) = 2.96, p < .008$. As shown in Table 2, a Tukey test revealed significant differences between the zero and six error conditions. There was no significant main effect for job status, $F(1, 147) = 1.50, p < .22$ nor was there a significant interaction between job status and speech errors, $F(2, 147) = .39, p < .89$. Post hoc analysis to test for gender differences proved nonsignificant.

Table 2
Main Effect of Speech Errors

# of Errors	n	Rating		
		Hire	Perform	Overall
0	68	3.97 ^a (.81)	3.89 ^a (.65)	3.91 ^a (.74)
2	26	3.76 ^{ab} (.86)	3.76 ^{ab} (.76)	3.69 ^{ab} (.83)
6	59	3.33 ^b (.75)	3.50 ^b (.62)	3.39 ^b (.67)

Because 30% of the original participants did not pass the manipulation check, a post hoc analysis was run including the entire sample of 217 students. Participants were assigned to one of the following three groups: grammatical errors made and perceived by the participant, grammatical errors made but not perceived by the participant, and grammatical errors not made and not perceived by the participant. As shown in Table 3, the ANOVA showed that if errors were perceived, favorability ratings were lower than if the errors were not perceived.

The data from this study suggest that reference checkers attend to speech errors made by a reference provider and that these errors will affect the perception of the job applicant's favorability. In other words, in the case of a reference call, the characteristics associated with using poor English are attributed not only to the reference giver, but somehow misattributed to the applicant, thereby falsely tainting the perception of the job applicant.

Table 3

Results of post hoc analysis

Condition	N	Mean	SD
Errors made and perceived	85	3.51 ^a	.80
Errors made but not perceived	60	3.99 ^b	.85
Errors not made, not perceived	68	3.92 ^b	.72

References

- Aamodt, M. G. (1996). *Applied Industrial/Organizational Psychology (second edition)*. Pacific Grove, CA: Brooks-Cole.
- Aamodt, M. G., Bryan, D. A., & Whitcomb, A. J. (1993). Predicting performance with letters of recommendation. *Public Personnel Management, 22(1)*, 81-90.
- Aamodt, M. G., Dwight, S., & Michals, J. (1994). *Reliability and validity of the trait approach to letters of recommendation*. Paper presented at the annual meeting of the International Personnel Management Association - Assessment Council, Charleston, SC.
- Ceci, S. J., & Peters, D. (1984). Letters of reference: A naturalistic study of the effects of confidentiality. *American Psychologist, 39(1)*, 29-31.
- Erisman, M. S., & Reilly, N. P. (1996). *The 360 degree performance evaluation: Unique or redundant information?* Paper presented at the eleventh annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.
- Kleinke, C. L. (1978). Perceived approbation in short, medium, and long letters of recommendation. *Perceptual and Motor Skills, 46(1)*, 119-122.
- Knouse, S. B. (1983). The letter of recommendation: Specificity and favorability of information. *Personnel Psychology, 36(2)*, 331-341.
- Mehrabian, A. (1965). Communication length as an index of communicator attitude. *Psychological Reports, 17*, 519-522.
- Miller, R. K., & Van Rybroek, G. J. (1988). Internship letters of recommendation: Where are the other 90%? *Professional Psychology: Research and Practice, 19(1)*, 115-117.
- Peres, S. H., & Garcia, J. R. (1962). Validity and dimensions of descriptive adjectives used in reference letters for engineering applicants. *Personnel Psychology, 15*, 279-286.
- Ralston, S. M., & Thamel, C. A. (1988). Effect of vividness of language on information value of reference letters and job applicants' recommendation. *Psychological Reports, 62(3)*, 867-870.
- Range, L. M., Menyhart, A., Walsh, M. L., Hardin, K. N., Ellis, J. B., & Craddick, R. (1991). Letters of

recommendation: Perspectives, recommendations, and ethics. *Professional Psychology: Research and Practice*, 22(5), 389-392.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, 35, 1-62.