Running head: OBSERVER-RATINGS USED FOR SELECTION

Cite: Zimmerman, R., Triana, M., Barrick, M. (2010) Predictive Criterion-Related Validity of Observer-Ratings of Personality and Job-Related Competencies using Multiple Raters and Multiple Performance Criteria. Human Performance, 23, 361-378.

doi: 10.1080/08959285.2010.501049.

Link to final paper at Taylor & Francis web site:

http://www.tandfonline.com/doi/abs/10.1080/08959285.2010.501049#.VavMxltRFjo

Predictive Criterion-Related Validity of Observer-Ratings of Personality and Job-Related Competencies using Multiple Raters and Multiple Performance Criteria

Ryan D. Zimmerman

Texas A&M University

María del Carmen Triana

University of Wisconsin – Madison

Murray R. Barrick

Texas A&M University

Authors' Note

Comments concerning this article should be addressed to Ryan Zimmerman, Mays Business School, Texas A&M University, 4221 TAMU, College Station, TX 77843-4221. Phone number (979) 845 – 3133. Electronic mail should be sent to rzimmerman@mays.tamu.edu.

Abstract

The purpose of this study was to assess the predictive validity of observer-ratings of personality and job-related competencies in a selection setting. Based on ratings from multiple raters of both the predictors and the criteria in a sample of MBA students, results indicated that observer ratings of conscientiousness, emotional stability, leadership, and interpersonal skills predicted work performance, team performance, and academic performance. For work performance and team performance, a composite of the four predictors had incremental predictive validity over general mental ability, even after controlling for how well the rater knew the ratee.

Predictive Criterion-Related Validity of Observer-Ratings of Personality and Job-Related Competencies using Multiple Raters and Multiple Performance Criteria

Despite decades of research and multiple meta-analytic reviews (Barrick, Mount, & Judge, 2001; Barrick & Mount 1991; Salgado, 1997; Zimmerman, 2008), there is still considerable debate as to the usefulness of personality in predicting outcomes important in organizational settings. Recently, two groups of selection researchers (Morgeson, Campion, Dipboye, Hollenbeck, Murphy, & Schmitt, 2007; Ones, Dilchert, Viswesvaran, & Judge, 2007) engaged in an exchange in the literature regarding this issue. While the two sets of authors disagreed on the utility of using personality testing in organizations, they did reach consensus on a few issues. One of these issues is the continued need for more research that goes beyond using self-reports from the applicants themselves to ratings obtained from relatives or acquaintances of the applicants (i.e., “observer-ratings”).

Consistent with this call for more research on observer-ratings of personality, there have only been a handful of published field studies utilizing observer-ratings of individuals’ personalities and job-related competencies to predict performance-related outcomes. However, the extant research available on observer-ratings is not without limitations. Specifically, many of the current observer-ratings studies utilize a concurrent validation design in a non-selection setting (e.g., Daniel, 1990; Mount, Barrick, & Strauss, 1994; Small & Diefendorff, 2006) and/or only collect ratings from one rater of the predictors, criteria, or both (e.g., McCarthy & Goffin, 2001; Taylor, Pajo, Cheung, & Stringfield, 2004). In addition, few, if any, examine the criterion-related validity of the observer-ratings in conjunction with multiple types of performance criteria or their incremental validity over and above any alternative selection methods. Therefore, this study will contribute to the selection literature by (a) utilizing observer-ratings of multiple personality and job-related predictors in a selection setting, (b) predicting multiple performance criteria, including team performance, academic performance and work performance, (c) using multiple raters of the predictors and performance criteria, and (d) assessing the incremental validity of the observer-rated predictors over an alternative selection tool (i.e., a test of general mental ability, GMA). If observer-ratings of personality and other job-related competencies collected in a selection-setting are effective in predicting future performance, this selection method will offer organizations a useful technique with which to hire future employees.

Observer-Ratings

Despite the fact that the use of observer-ratings helped to provide the foundation to five-factor model (FFM) personality research (Tupes & Christal, 1961) and is the core methodology used in other areas of human resource management research (e.g., multisource feedback and the employment interview), there has been little attention paid to evaluating the efficacy of observer-ratings of personality in employee selection settings. Two major reasons why skeptics doubt the usefulness of personality assessment in organizational research, particularly in employee selection, include their concern over social desirability affecting applicants' self-reports of their personality traits, as well as their frustration over self-reports having low criterion-related validities with important organizational outcomes (McCrae & Weiss, 2007; Morgeson et al., 2007). In order to help address these issues, critics recommend greater research attention be paid to observer-ratings of personality traits.

An additional reason that observer-ratings of personality traits and other job-related competencies warrant greater attention in the research literature is the frequency with which organizations utilize information collected from those acquainted with the applicants (e.g., either personal references and/or previous employers) of individual differences variables during the hiring process. As shown by the results of a survey conducted by the Society for Human Resource Management (2005), the vast majority (96%) of organizations use some sort of reference check procedure, while only 30% use personality tests in hiring. This fact implies that academics have overlooked conducting research on a selection method that practitioners heavily rely upon (for exceptions, see McCarthy & Goffin, 2001; Taylor et al., 2004).

Past research on observer-ratings of personality has consistently shown that such ratings can be valuable (McCrae & Weiss, 2007). Observer-ratings represent the individuals' reputations, which is based on their public actions and behaviors that have been exhibited over time. In fact, Hogan (1996) is critical of the heavy reliance on self-reports in personality research. First, he argues that the trait-based items commonly used to evaluate personality are more appropriate for how observers describe individuals' personalities, but not how individuals naturally describe themselves. Second, he notes that self-reports of personality are best considered as "self-presentations" that are malleable self-evaluations and that any "consistencies in the pattern of a person's item endorsement reflect consistent styles of self-presentation rather than underlying traits" (p. 175). However, Hogan argues that observer-ratings are stable judgments of the ratees' reputations, which have been established based on consistent and publicly-observable past behaviors. As past behaviors are the best predictor of future behaviors (Wernimont & Campbell, 1968), observer ratings of reputation-based personality should therefore be good predictors of future performance criteria that are composed of multiple instances of the ratees' behaviors.

There is empirical support for Hogan's arguments. In relation to the distinction between self-ratings and observer-ratings, Connolly, Kavanagh, and Viswesvaran (2007) used meta-analytic estimates to show that while the two types of ratings of personality are strongly related, the two are not perfectly correlated even after correcting for measurement error based on internal consistency estimates of reliability. This indicates that the differences between self- and observer-ratings are substantive and not merely a statistical artifact. There is also empirical support for the hypothesis that observer-ratings of past behaviors, compared to self-ratings of a malleable self-image, would be the most stable over time and would therefore be the best predictors of performance criteria related to such behaviors. In non-selection settings, observer-ratings have been shown to have more incremental validity over self-ratings than vice versa (Bratko, Chamorro-Premuzic, & Saks, 2006; Kolar, Funder, & Colvin, 1996; Mount et al., 1994; Small & Diefendorff, 2006). Because of their potential for greater criterion-related validity, observer-ratings of personality and other job-related competencies should be regarded as an important underresearched selection tool that warrant greater attention in the academic literature.

Hogan's (1996) distinction between self-ratings, as internal self-concept, and observer-ratings, as past public behaviors, may also aid in the identification of a useful method to reduce socially-desirable responding. Whereas self-reports may be influenced by unconscious inflation (i.e., self-deception) due to internal aspirations or self-image, as well as conscious inflation (i.e., impression management) of personality test scores in order to obtain a desired outcome (Paulhus, 1984), observer-ratings would be less affected by both factors. First, by definition, observers would not be affected by the focal individuals' self-deception. Second, observers should be less likely to exaggerate their evaluations of the focal individuals as they likely have far less to gain by inflating the test scores compared to the applicants.

However, in a selection setting, applicant-chosen raters may be inclined to elevate their ratings in hopes of increasing the chances that the applicants would be selected. That is, applicants who select their references are apt to choose raters who are likely give the applicants' the highest scores. Therefore, there is a question as to whether there will be enough variance in applicants' scores received from applicant-selected observers to allow for meaningful prediction of performance criteria. Fortunately, this is an empirical question that can be best answered by utilizing a research design such as the one used in this study.

Finally, as noted by McCrae and Weiss (2007), there are advantages to collecting and aggregating ratings from multiple raters. First, utilizing multiple raters will increase the inter-rater reliability of the observers' ratings, and thus improve the accuracy of the assessments. Second, as discussed in the employment interview literature, increasing reliability will, all else equal, increase the criterion-related validity of the ratings (Schmidt & Zimmerman, 2004). However, as noted by McCrae and Costa (1989) and Schmidt and Zimmerman (2004), when elevating the number of raters above three, the corresponding increase in inter-rater reliability begins to plateau. Therefore, as part of this study we collect the predictor ratings from three observers. In addition, as the benefits of utilizing multiple raters also applies to performance criteria (Viswesvaran, Ones, & Schmidt, 1996), we also collect performance ratings from three (to four) raters. We see this as one of the major strengths of the study, as we are aware of no other studies that have used multiple raters of both predictors and multiple performance criteria in a selection setting to validate the criterion-related validity of observer-ratings of personality and other job-related competencies.

Constructs Assessed

Based on a comprehensive job analysis appropriate for this study's setting (see Method section), four non-cognitive predictors of performance were identified for inclusion in a reference checklist (Aamodt, 2007, p. 144) that was used as a letter of recommendation to collect the observers' ratings for applicants to an MBA program. Specifically, subject matter experts

(SMEs), including the administrators of the MBA program and the researchers conducting this study, determined that the constructs of conscientiousness, emotional stability, leadership, and interpersonal skills were critical to the success of the students in the MBA program, as well as in their professional careers. Performance would be measured by three criteria: academic performance, team performance, and work performance. Academic performance reflects how well the participant did in their core courses during the first semester of the MBA program. Team performance represents the ratings that the participants received from their project team members. Finally, work performance represents the employees' performance on their jobs, as rated by supervisors and coworkers.

The personality traits of conscientiousness and emotional stability were selected for inclusion in the reference checklist because they consistently have been found to predict a wide variety of performance criteria, including the three measured in this study. A meta-analysis by Barrick et al. (2001) of self-reports of personality determined that conscientiousness and emotional stability have true score correlations of .24 and .15 (respectively) with work performance, and .27 and .22 with team performance. In addition, several primary studies (Bratko et al., 2006; Chamorro-Premuzic & Furnham, 2003; Lounsbury, Huffstetler, Leong, & Gibson, 2005; Paunonen & Ashton, 2001) have found self-reports of conscientiousness ($\bar{r}$ = .28, range: .14 - .38) and emotional stability ($\bar{r}$ = .13, range: -.13 - .30) to predict academic performance. Bratko et al. (2006) also examined peer-ratings of personality and identified peer-ratings of conscientiousness to be a strong predictor of academic performance (*r* = .54); however, although peer-ratings of emotional stability were positively correlated with academic performance (.05), the result was not statistically significant.

Similarly, leadership and interpersonal skills were selected because they are also likely to relate to the three performance criteria. A meta-analysis by Huffcutt, Conway, Roth, and Stone (2001) found leadership and interpersonal skills (as measured through the interview) to have true score correlations of .47 and .39 with work performance. In another meta-analysis, Judge, Piccolo, and Ilies (2004) estimated the true score correlation between observer-rated leadership and work performance to be .25 and between leadership and team performance to be .29. While meta-analytic results are not available for the relationship between interpersonal skills and team performance, a recent primary study by Morgeson, Reider, and Campion (2005) found a significant relationship between social skills and team performance (r = .28), even when controlling for personality and teamwork knowledge (β = .18).

Finally, because an individual's performance in an MBA program is almost always heavily dependent on grades based on team assignments (Baldwin, Bedell, & Johnson, 1997), we also believe leadership and interpersonal skills will influence academic performance. Previous research has found that academic performance is predicted by both leadership (r = .46, Mohammed, Mathieu, & Bartlett, 2002) and interpersonal skills (r = .23, Gilman & Anderman, 2006). In sum, we expect:

Hypothesis 1: Observer-ratings of Conscientiousness (1a), Emotional Stability (1b), Leadership (1c), and Interpersonal Skills (1d) will positively predict academic performance.

Hypothesis 2: Observer-ratings of Conscientiousness (2a), Emotional Stability (2b), Leadership (2c), and Interpersonal Skills (2d) will positively predict team performance.

Hypothesis 3: Observer-ratings of Conscientiousness (3a), Emotional Stability (3b), Leadership (3c), and Interpersonal Skills (3d) will positively predict work performance.

Incremental Criterion-Related Validity Beyond General Mental Ability

In addition to reference checks and personality testing, another selection method that is used by organizations is an assessment of GMA. As such, it is an alternative predictor with which to compare the observer-rated non-cognitive predictors. Therefore, in order to assess the usefulness of observer-ratings of conscientiousness, emotional stability, leadership, and interpersonal skills in an organization's selection process, we will evaluate the incremental criterion-related validity of these constructs over and above GMA.

We expect the four observer-rated predictors to have incremental validity over GMA based on previous research which shows that the predictors included this study are not strongly related to GMA. Prior research found that conscientiousness and emotional stability have weak relationships with GMA, with correlations ranging between .00 to .14 when personality is self-reported (Bratko et al., 2006; Busato, Prins, Elshout, & Hamaker, 2000; Goff & Ackerman, 1992; Ones, 1993) and between .00 to .15 when personality is rated by observers (Bratko et al., 2006; Huffcutt et al., 2001). Similarly, observer-ratings of leadership and interpersonal skills have demonstrated weak relationships (.00 to .19) with GMA (Huffcutt et al., 2001; Judge, Colbert, & Ilies, 2004). Because the four observer-rated predictors included in this study have been shown to relate more strongly to the performance criteria than the predictors do with GMA, we believe they will explain a significant amount of variance in the performance criteria above GMA. Therefore, we conclude:

Hypothesis 4: Observer-ratings of the four observer-rated predictors will account for variance in the performance criteria beyond that accounted for by GMA.

Method

Participants and Procedures

The sample consisted of applicants to a graduate business (MBA) program at a large university in the Midwest region of the United States. There were a total of 712 applicants. Of these applicants, 141 enrolled. The typical participant was white (60%), male (75%), approximately 30 years old, with 6.41 years of full-time work experience (*SD* = 3.04). As part of the MBA program's selection process, each applicant had to have three individuals submit a reference checklist/letter of reference on the applicant's behalf. Both the applicants and their referents knew that the referents' ratings would be used as part of the selection process. The letter of reference form stated that the information provided would be used for research purposes. Additional informed consent was obtained from students during the MBA program to use the predictors and criteria collected. During the first semester of their coursework, all participants had their academic performance evaluated in each of their courses based on a standard grade point average scale (0 - 4.0).

Additionally, in their management course during their second semester, all participants formed project teams with typically five members in each group. As part of the course, the teams had to evaluate four business cases and write a report on each, as well as present their solution for one of the cases. At the end of the course, all of the team members evaluated each other on their technical proficiency in completing the case assignments and contextual performance within the team. This data was collected over three years.

For one year, MBAs that started the program together participated in a developmental program. The developmental program was only offered to this one cohort as part of their orientation week before the start of their first semester of coursework. Part of the program involved asking current coworkers (peers and supervisors) to evaluate their work performance through a confidential online survey system. These participants received ratings from an average of nearly four raters. Complete academic data were available for 127 of the 141 participants and 44 of these had work performance data.

Measures

Observer Ratings

To collect the observer ratings, we utilized a reference checklist that included scales to predict success in the academic program, as well as success in a variety of future work assignments. Hence, we focused on predictors that would relate to three broad performance dimensions: teamwork, academic success, and overall work performance. Three SMEs and two job analysts individually generated a list of potential KSAs. The SMEs had graduate degrees (1 PhD, 2 MBAs) with a mean of eleven years (*SD* = 3.4) of administering the MBA program and selecting students for the program. The two job analysts also had graduate degrees (both were PhDs) with a mean of twelve years of experience in developing and implementing selection tests/procedures.

The four broad competencies used in the reference checklist (conscientiousness, emotional stability, leadership, and interpersonal skills) were generated as an outcome of this process. Measures for the four selected constructs are described below. To evaluate the feasibility of using the finished reference checklist with the full-time MBA program, the instrument was utilized with one group of students who were applying to the evening MBA program for mid-year enrollment. Applicant and reference reactions were positive to the new form. Hence, the new form was adopted for use with full-time MBAs for the following fall enrollment.

The reference checklist containing the observer ratings consisted of four scales: conscientiousness, emotional stability, interpersonal skills, and leadership. Each scale used a five-point Likert rating format (from 1 = *strongly disagree* to 5 = *strongly agree*). To create each applicant's score on the overall checklist, we calculated the applicant's grand mean across the four scales (i.e., we averaged the applicant's mean response on each of the scales).

Conscientiousness was measured with nine items from the Wonderlic Productivity Index® (WPI; Barrick, Mount, & Wonderlic, 2006). The WPI was designed to measure personality traits based on the FFM of personality. Items were related to applicants' reliability, persistence, work ethic, and orderliness. Example items include "Before beginning his or her work, this person likes to plan and organize it" and "This person is very thorough in any work he or she does". The scale had a coefficient alpha of .71 (Cronbach, 1951).

Emotional Stability was measured with eight items from the WPI and had a coefficient alpha of .68. Items measured the applicants' tendencies to be calm, secure, and well-adjusted. Sample items include "This person tends to be very secure with himself or herself" and "At times this person spends too much time worrying about unimportant things".

Leadership was measured with seven items from the Global Transformational Leadership scale by Carless, Wearing, and Mann (2000). Sample items are "This person gives encouragement and recognition to others", "This person instills pride and respect in others and inspires them by being highly competent", and "This person encourages thinking about problems in new ways and questions assumptions". The scale had a coefficient alpha of .87.

Interpersonal Skills was measured with eight items developed for this study. The items include "This person seldom offends other people", "This person works well with all types of people", "I have never seen another person who can interact as well with others as this person", "All types of people enjoy interacting with this person", "This person could be considered a 'people-person'", "Just like everyone, there are some people that would not get along with this person”, “Sometimes this person will argue with others' ideas without necessarily having an alternative”, and “This person always thoughtfully considers others' suggestions”. The scale had a coefficient alpha of .76. The composite of all four scales had a coefficient alpha of .80.

Academic Performance

Academic performance consists of the first semester grade point average (GPA) across five courses (marketing, finance, accounting, statistics, and economics). GPA followed a standard four point scale with a course grade of an ‘A’ corresponding to four points, a ‘B’ equal to three points, etc. In order to ensure that differences in GPA were not influenced by differences in courses taken or instructors, only the first semester’s GPA was used in the analyses. All MBAs had the same courses with the same instructors for the first semester of the program.

Team Performance

Performance of the individual on the team projects was rated by team members on two scales: technical proficiency and contextual performance. Each scale used a five-point rating format (from 1 = *never* to 5 = *always*). Technical proficiency was rated by three items: “Completed assigned work thoroughly and accurately”, “Communicated skillfully in written and oral communications”, and “Understood principles and ideas taught in the class”. The scale had a coefficient alpha of .83. Contextual performance was measured with eight items. The items included “Showed up on time for team meetings”, “Came prepared for team meetings”, “Met team deadlines for completing work”, “Completed his or her fair share of team workload”, “Actively participated in discussions on team project-related issues”, “Went above and beyond requirements by volunteering for extra work”, “Helped the team set the agenda for meetings”, and “Fostered trust, involvement, and cooperation among the team”. The scale had a coefficient alpha of .93. The combined eleven-item scale had a coefficient alpha of .95. Scores on the two scales were averaged to arrive at the individual’s total *Team Performance*. As the team projects were part of the management course taken by the MBAs during their second semester in the program, academic performance and team performance are independent criteria.

Work Performance

Work performance was gathered from recent supervisors and coworkers after the initial checklist scales were completed, but before the incoming participants started their formal coursework for the MBA program. The individual’s performance in the work place was measured with four dimensions. Each dimension was rated on a five-point scale from 1 = *Unsatisfactory* to 5 = *Far Exceeds Expectations*. The first dimension included in the scale was Work Effort, defined as how well the employee focuses on maintaining strong performance, contributes extra effort as necessary to complete tasks successfully and accomplish work goals, and persists despite obstacles and occasional set backs. The second dimension was Initiative, defined as how well the individual goes above and beyond job requirements by volunteering for extra work activities that are not part of the job, and suggests organizational improvements. The third dimension was Resilience, defined as how well the person copes with stress effectively without allowing it to interfere with performance or disrupt working relationships. The fourth dimension was Adaptability, defined as how well the employee adapts behaviors and performs job requirements in ambiguous or changing conditions, interacts flexibly with coworkers and customers, and adjusts to changing job demands and situations. The scale had a coefficient alpha of .81.

General Mental Ability

Scores on the Graduate Management Admissions Test were collected as part of the admissions procedures. As the standardized test measures verbal and quantitative ability, two components of most assessments of general intelligence, it is used as a measure of GMA similar to selection tests that could be used during an organization's hiring process. The scale used ranges from 200 to 800.

Familiarity with Ratee

Finally, we measured the extent to which the rater filling out the reference checklist was familiar with the ratee. Although we did not have any hypotheses about familiarity, we collected this variable because previous research has shown that a rater's level of familiarity with the ratee can enhance the ratings they provide (Judge & Ferris, 1993; McCrae & Weiss, 2007; Taylor et al., 2004). Therefore, this variable was collected simply for use in post-hoc analyses to show whether the observer-ratings' predictive validity will hold even when controlling for a rater's familiarity with the ratee. This construct was measured with one item: "How well do you believe you know this person at work or at school?" The item had a response format ranging from 1 = *not at all* to 5 = *extremely well*.

Aggregation of Ratings

As there were multiple raters of the four predictors, team performance, and work performance, the average of the raters' scores was used to create participants' total scores on each of the measures. In order to justify this aggregation, the intraclass correlation coefficient(2) was computed for each scale (Bliese, 2000; James, 1982). The ICC(2)s ranged from .63 to .73, which indicates adequate agreement to aggregate the ratings across the multiple raters (Atwater, Ostroff, Yammarino, & Fleenor, 1998; Kristof-Brown & Stevens, 2001).

Results

Table 1 reports the means, standard deviations, inter-correlations, and coefficients alpha of the variables. Correlations marked with an asterisk are significant at the .05 level. We conducted regression analyses to test Hypotheses 1 to 4. Results for all three regression analyses are shown in Table 2. All of the correlations between the four predictors and the three measures of performance are positive and most of them are significant.

Academic performance was significantly related to conscientiousness, leadership, and interpersonal skills with correlations of .16, .19, and .22, respectively. These results support Hypotheses 1a, 1c, and 1d. Although the correlation is positive as posited, the result for Hypothesis 1b (academic performance) was not significant. Team performance had significant correlations with conscientiousness (.27) and emotional stability (.30), which supports Hypotheses 2a and 2b. While the leadership-team performance and interpersonal skills-team performance relationships were positive, neither was significant. Therefore, Hypotheses 2c and 2d were not supported. Finally, work performance was positively and significantly related to each of the predictors, with correlations of .35, .26, .51, and .29 for conscientiousness, emotional stability, leadership, and interpersonal skills, respectively. These results support Hypotheses 3a, 3b, 3c, and 3d.

We then conducted a regression analysis in order to assess how strongly the four observer-rated predictors related to our three measures of job performance. Model 1 of Table 2 presents these results. The set of four predictors significantly and positively predicts academic performance (R = .25, p < .05), team performance (R = .36, p < .05), and work performance (R = .52, p < .05). In terms of unique variance explained by each of the four predictors, conscientiousness and emotional stability explained unique variance in team performance (βs = .26 and .23, respectively), while leadership explained unique variance in work performance (β = .58). While none of the predictors explained unique variance in academic performance, their shared variance was significant. These results further support Hypotheses 1, 2, and 3.

Hypothesis 4 predicted that the observer-ratings of conscientiousness, emotional stability, leadership, and interpersonal skills would account for variance in the performance criteria beyond GMA. In order to test this hypothesis, we conducted a hierarchical multiple regression with GMA in Step 1 and the four observer-rated non-cognitive predictors in Step 2. Model 2 of Table 2 shows these results. The set of four observer-rated predictors significantly predict team performance above GMA ($\Delta R^2 = .13$, $p < .05$). The four predictors also predict work performance above GMA ($\Delta R^2 = .27$, $p < .05$). However, the observer-rated variables are not a significant predictor of academic performance beyond GMA. Although observer-ratings of the four predictors explained 5% of the variance in academic performance beyond GMA, this increase was not significant. Overall, these results provide support for Hypothesis 4 except for academic performance. As a set, the observer-ratings of conscientiousness, emotional stability, leadership, and interpersonal skills do predict a significant amount of variance beyond GMA for both team performance and work performance.

We conducted an additional post-hoc analysis where we assessed the predictive validity of the observer-ratings beyond the rater's familiarity with the ratee. Although we did not propose any hypotheses about this, we conducted this post hoc analysis to ascertain the degree to which the rater's level of familiarity with the ratee affected the results of the regressions of the performance criteria on the set of four observer-rated predictors (Judge & Ferris, 1993; McCrae & Weiss, 2007; Taylor et al., 2004). Therefore, we conducted a hierarchical multiple regression with GMA in step 1, the rater's reported familiarity with the ratee in Step 2, and the four predictors in Step 3. As Model 3 in Table 2 shows, the results are stable even after controlling for the effects of the rater's familiarity with the ratee. The set of observer-rated predictors still significantly predicts both team performance and work performance beyond the GMA. The beta coefficients for each of the four predictors in Model 3 are very similar to those in Model 2. This provides additional support for Hypothesis 4.

Discussion

The four observer-rated predictors examined (conscientiousness, emotional stability, leadership, and interpersonal skills) were significantly related to the three performance criteria: academic performance, team performance, and work performance. Further, this set of four predictors had significant incremental validity over GMA for team performance and work performance, even after controlling for rater familiarity with the applicant.

As suggested by Hogan's (1996) socio-analytic theory of personality, observer-ratings of personality-related constructs were good predictors of individuals' future performance, whether measured as work, team, or academic performance. It is also important to note that the performance criteria were collected significantly after the checklist scales were completed. Typically (except in unusual cases of deferred enrollments), work performance was collected three to six months, academic performance was collected six to nine months, and team performance was collected ten to twelve months after the predictors. The diversity of settings and behaviors that were reflected in the performance criteria also confirm Hogan's (1996) assertion that observer-ratings based on an aggregation of ratees' past behaviors (i.e., their reputations) would be predictive of theoretically-related future behaviors, even across contexts. Additionally, these criterion-related validities were both statistically significant and practically meaningful despite the fact that applicants chose their raters, presumably based on who would give them the most favorable evaluations. That is, although the raters may have been biased in favor of the applicants, there was still enough variance to allow the checklist scales to be useful predictors.

We now turn to a discussion comparing the findings from the present study to other findings in the selection literature based on constructs assessed and assessment method. We acknowledge that many of the results we use for comparison purposes below are based on meta-analyses of studies in a variety of settings, whereas our study only includes one sample of MBA students. We make no claims of being able to definitively say which method is better based on just one study. However, as others studying selection have done in the past (e.g., Schmidt & Hunter, 1998, for example), we discuss our findings relative to findings from other studies because it is helpful to compare the predictive validity of various selection methods in order to understand the efficacy of different methods.

Comparison of Findings Based on Constructs Assessed

Based on the findings of this study, observer-ratings of conscientiousness and emotional stability may be stronger predictors of both work and team performance compared to self-ratings of these two traits. Observer-ratings of conscientiousness were correlated .35 and .27 with work performance and team performance, respectively; whereas Barrick et al. (2001) meta-analytically estimated the average observed relationships across settings between self-ratings of conscientiousness and work and team performance to be .12 and .15. For emotional stability, observer-ratings were correlated .26 and .30 with work and team performance, with the average observed relationship utilizing self-ratings estimated at .09 and .13 (Barrick et al., 2001). Additionally, the criterion-related validities found in this study for conscientiousness and emotional stability are substantially larger than those found in the meta-analysis by Conway, Lombardo, and Sanders (2001) for the relationship between self-ratings of personality, and peer and subordinate ratings of job performance (dependability $\bar{r}$ = .11 with peer ratings of job performance; .01 with subordinate rating of performance; adjustment $\bar{r}$ = .09 and .06, respectively).

The criterion-related validities of the four observer-rated predictors are similar to those found for the same constructs when they are measured during the employment interview. The meta-analysis by Huffcutt et al. (2001) that examined the criterion-related validities of observer-rated constructs measured in the employment interview indicated that conscientiousness, emotional stability, leadership, and interpersonal skills had average observed correlations with overall job performance of .18, .26, .26, and .21, respectively. These criterion-related validities are similar to those found in this study where the constructs are measured through applicant-selected observers instead of organization-selected interviewers. Specifically, in this study, conscientiousness, emotional stability, leadership, and interpersonal skills had average correlations across all three performance criteria (academic performance, team performance and work performance) of .26, .19, .25, and .21, respectively.

However, our criterion-related validities for leadership are stronger than those of Zimmerman, Mount, and Goff (2008), who found that peer and subordinate multisource feedback ratings of leadership had an average correlation of .15 with overall performance when the leadership ratings were made for developmental purposes. Conversely, our findings for leadership are the same as Zimmerman et al.'s criterion-related validities ($\bar{r}$ = .25) when the leadership ratings were made for administrative decisions (e.g., pay raises or promotions). Finally, while the correlation between leadership and work performance (.51) was stronger than the average observed-correlation (.19) found by Judge et al. (2004), the relationship between leadership and team performance (.05) was much weaker (.23; Judge et al., 2004). However, as the subjects in this study were not "true" managers and would have to emerge as informal leaders, the latter finding is not entirely surprising. In sum, the criterion-related validities found in this study are usually at least similar, if not stronger, than those found by other researchers investigating the criterion-related validity of observer ratings of similar constructs in non-selection settings.

In relation to the previous paragraph, it should be emphasized that although previous research has examined the validity of observer-ratings of personality and job-related competencies, these studies have not typically been conducted in selection settings (Judge, Bono, & Locke, 2000; Mount et al., 1994). Usually, these studies would ask current employees to have a significant other or coworker rate the employee’s personality. These ratings were then correlated with job satisfaction or job performance. The majority of studies that have examined the criterion-related validity of observer ratings of personality and other job-related competencies in a selection setting focused on interviewer ratings (see Huffcutt et al., 2001 for a review). Thus, the “observer” in the interview setting has a duty to the organization to evaluate the ratee accurately, but has had only limited contact with the individual. However, in this study, the applicant chose the raters, all of whom would likely have greater feelings of obligation to and bias in favor of the applicant. It is also true that they have known the individual for much longer than an interviewer at an organization would have. Therefore, an additional contribution of this study is to the understanding of the criterion-related validity of observer ratings of personality and job-related competencies when conducted in a selection setting and when the raters are chosen by the applicant. This contribution is responsive to recent calls by prominent personnel selection researchers who suggested that more research is needed on non-self reports of such constructs in selection contexts (Morgeson et al., 2007; Ones et al., 2007).

Comparison of Findings Based on Assessment Method

As we used a reference checklist as the selection method to collect the observer-ratings, some discussion of this method is warranted. In his early research using observer-ratings of personality to construct the FFM, Norman (1963) noted that the information obtained from observer-ratings of personality are akin to the information found in letters of reference. Further, because of the design of the reference checklist used in this study, it could be regarded as a “structured” reference check commonly used by organizations. That is, as outlined by previous researchers (Campion, Palmer, & Campion, 1997; Chapman & Zweig, 2005) regarding the process needed to standardize the employment interview, the reference checklist used in this study meets those same requirements. Specifically, the process used to create the reference checklist used in this study included a) using the results of a job analysis to select job-relevant constructs to measure; b) asking the same, specific questions of all referents and avoiding unstructured, open-ended questions; c) using a longer reference by asking more content-valid questions; d) relying on multiple raters to obtain a broader sample of ratees’ behaviors; and e) using a set scoring key to avoid shifting standards (Campion et al., 1997; Chapman & Zweig, 2005). Therefore, it is useful to make some comparisons between our observer-rated reference checklist and other selection methods.

The average criterion-related validity of the composite of the four observer-rated predictors found in this study ($\bar{r}$ = .28) is larger than the meta-analytic estimate ($\bar{r}$ = .18) found by Reilly and Chao (1982) for the criterion-related validity of reference checks. After correcting for range restriction and unreliability in the criterion, the average corrected correlation of our composite ($\bar{\rho}$ = .42) is also larger than the corrected estimate ($\bar{\rho}$ = .26) found by Hunter and Hunter (1984) for reference checks. However, it should be noted that the studies included in the two aforementioned meta-analyses nearly all used unstructured reference checks. In comparisons to the employment interview, the relative increase in validity of our structured reference checklist over unstructured reference checks (56%; $\bar{r}$ = .28 and .18, respectively) is larger than the increase in the criterion-related validity of structured interviews over unstructured interviews (34%; ρ = .51 and .38; McDaniel, Whetzel, Schmidt, & Maurer, 1994).

The average corrected validity of the composite of the four observer-rated predictors ($\bar{\rho}$ = .42) is comparable to the corrected validities of integrity tests (.41), unstructured interviews (.38), assessment centers (.37), and biodata measures (.35) (Schmidt & Hunter, 1998). Our results suggest that observer-rated structured reference checklists can offer organizations an additional selection method with high criterion-related validity. As most organizations utilize some sort of reference check already, structuring the reference check method would substantially increase the utility of this method beyond the unstructured reference checks typically used. Finally, structured reference checklists may provide greater legal defensibility since the same questions are asked of all applicants using a standardized response key, without the possibility of personal biases from the organization's recruiter affecting how open-ended responses are evaluated.

As we are collecting multiple observer-ratings in a selection setting, it is also important to compare the inter-rater reliability of such ratings to inter-rater reliabilities from other selection methods, as well as observer-ratings collected in non-selection settings. These comparisons will provide an indicator of the degree of consensus between raters when assessing ratees' behaviors for different purposes. By comparison, the inter-rater reliability of the composite of four predictors used in this study (.45) is nearly double that of unstructured reference checks (.23; Aamodt & Williams, 2005). This increase in inter-rater reliability is of similar magnitude to that of structured interviews compared to unstructured interviews (.67 to .34; Conway, Jako, &

Goodman, 1995). However, the inter-rater reliability of the measures collected in our selection context is less than the average inter-rater reliability of observer-rated personality measures collected in non-selection settings (.59, Connolly et al., 2007).

Limitations and Future Research

This study has some limitations that should be addressed in future research. First, the sample size for the work performance criterion is fairly small. However, the relationships between each of the four observer-rated predictors with work performance ratings collected from the participants' supervisors were still significant.

Second, although the incremental validity of the observer-ratings was established over GMA in this setting, two limitations arise from this finding. We note that although incremental validity of GMA was established in this setting, the magnitude of the relationship between GMA, and team and work performance, was not significant. As range restriction on GMA could be one reason for the small magnitude of these correlations, future research should establish incremental validity over GMA in a sample comprised of a less skewed distribution of intellectual ability. In addition, future research should examine their incremental validity over other commonly used selection methods. In particular, the incremental validity over the interview and self-reports of personality, leadership, and interpersonal skills should be established. Given that the MBA admissions process did not utilize either self-reports of the four predictors or interviews, it was unfeasible to establish incremental validity over such selection methods in this setting. Incremental validity over alternative ways of measuring personality and other job-related competencies would be a strong test of the utility of using observer-ratings of the constructs during the selection process. Further, depending on the type of job for which the organization is hiring, observer-ratings of other relevant constructs (e.g., other personality traits or job-related skills) could be included in future research.

Third, as the participants in this study were MBA students, there is some question as to whether the results will generalize to other contexts. Specifically, if academic and workplace settings place different demands and constraints on individuals, then the absolute or relative magnitude of each of the four predictor's relationships with the performance criteria may differ. For example, given that weak situations are typically regarded as allowing for greater variance in personality-driven behaviors (Barrick et al., 2001), if an academic context presents a weaker (or stronger) situation than a work context, than the magnitude of the personality-related correlations found in a work context could be weaker (or stronger). This issue also places a limitation on the comparisons of the results from this study with prior meta-analytic work, as prior meta-analyses have aggregated results across (typically workplace) settings. However, there are reasons to believe the results will generalize. Work performance was gathered from coworkers and peers after the initial predictors were completed, but before the participants started the coursework for their MBA program. The team performance ratings were based on how well the participants interacted with their teams and how well they completed team-oriented tasks, including understanding task-related knowledge, effective written and oral communication skills, and completing work thoroughly and accurately. The academic performance criterion reflects the acquisition of job knowledge and accurately completing job content-related assignments. Each of the criteria reflects important work-related behaviors. Furthermore, across these diverse performance criteria, whether obtained in an academic environment or in an actual work setting, the magnitudes of the predictive validities were remarkably similar. In order to alleviate the concern over generalizability, as well as establish incremental validity over self-reports of personality, a strong future research design would be to replicate and extend the findings in this study in a work setting that also used self-reports of personality during the selection process.

Fourth, raters were selected by the ratees, and as such may view themselves as an advocate for the applicant, and may fear the possibility of lawsuits, all of which can distort ratings by ignoring real differences in ratees and giving applicants artificially high ratings. This problem is not unique in our setting though, as this is true in almost every organization’s selection process, as applicants choose which individuals they list as references. More importantly, we contend that focusing on specific job-relevant predictors rather than requesting overall, global evaluations, should elicit more responses using a broader range of scores, thereby reducing leniency. In this setting, we did find that the four observer-rated predictors still had predictive validity for three different performance criteria, even when controlling for how familiar the rater was with the ratee.

A final limitation regarding the practical significance of our study is that some organizations may be reluctant to rely too heavily on references because they are concerned over asking for or providing information about individuals beyond dates of employment, job titles, and other such objective facts. However, recent developments are changing this perspective. First, 40 states have passed laws protecting employers who give detailed reference information about former employees, as long as that information was provided in good faith (Gatewood, Feild, & Barrick, 2008). Second, a survey of organizations revealed that few (2%) have had any legal issues regarding defamation of former employees, while slightly more (4%) have had legal issues due to negligent hiring or not providing adequate warning regarding the threat posed by a former employee (SHRM, 2005), both of which may be avoided by actually conducting more thorough reference checks. Third, based on the results of the same survey, organizations are much more likely to provide detailed reference information if the former employee has signed a waiver that limited their right to take legal action based on the information provided by the former employer (SHRM, 2005). Taken together, these recent developments suggest that a structured letter of reference does have practical significance for organizations.

Conclusion

This study is a response to recent calls for more research on the predictive criterion-related validity of personality and other job-related competencies. Based on the results of this study, the answer to these calls is that observer-ratings of conscientiousness, emotional stability, leadership, and interpersonal skills can be good predictors of team, academic, and work performance. In addition, for work performance and team performance, the set of four predictors had incremental predictive validity over general mental ability. As the vast majority of organizations use reference checks during their selection process, the use of standardized scales of personality and job-related competencies offers a way for them to increase the reliability and relevance of the information collected from acquaintances of the applicant. We hope this research allows organizations to increase the efficacy of their selection procedures, as well as continues to inform the academic field that indicators of personality are useful predictors of important organizational outcomes.

References

Aamodt, M. G. (2007). Industrial/Organizational Psychology: An Applied Approach (5th edition). Pacific Grove, CA: Wadsworth Publishing.

Aamodt, M. G., & Williams, F. (2005). *Reliability, validity, and adverse impact of references and letters of recommendation*. Paper presented at the 20th Conference of the Society for Industrial-Organizational Psychology, Los Angeles, CA.

Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other agreement: Does it really matter? *Personnel Psychology, 51*, 577-598.

Baldwin, T. T., Bedell, M. D., & Johnson, J. L. (1997). The social fabric of a team-based M.B.A. program: Network effects on student satisfaction and performance. *Academy of Management Journal, 40*, 1369-1397.

Barrick, M.R. & Mount, M.K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Barrick, M.R., Mount, M.K., & Judge, T.A. (2001). Personality and performance at the beginning if the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*, 9-29.

Barrick, M.R., Mount, M.K., & Wonderlic, Inc. (2006). *Wonderlic Productivity Index*. Wonderlic, Inc., Libertyville, IL.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349-381). San Francisco: Jossey-Bass.

Bratko, D., Chamorro-Premuzic, T., & Saks, Z. (2006). Personality and school performance: Incremental validity of self- and peer-ratings over intelligence. *Personality and Individual Differences, 41*, 131-142.

Busato, V.V., Prins, F.J., Elshout, J.J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality & Individual Differences, 29*, 1057-1068.

Carless, S. A., Wearing, A. J., & Mann, L. (2000). A short measure of transformational leadership. *Journal of Business and Psychology, 14*, 389-405.

Campion, M.A., Palmer, D.K., & Campion, J.E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50*, 655-702.

Chamorro-Premuzic, T., & Furnham, A. (2003). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality, 37*, 319–338.

Chapman, D.S., & Zweig, D.I. (2005). Developing a nomological network for interview structure: Antecedents and consequences of the structured selection interview. *Personnel Psychology, 58*, 673–702.

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment, 15*, 110-117.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565-579.

Conway, J. M., Lombardo, K., & Sanders, K. C. (2001). A meta-analysis of incremental validity and nomological networks for subordinate and peer ratings. *Human Performance, 14*, 267-303.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16,* 297-334.

Daniel, D. (1990). Validity of a standardized reference checklist. *Applied H.R.M. Research, 1,* 51–66.

Gatewood, R.D., Feild, H.S., & Barrick, M. (2008). *Human Resource Selection.* Sixth edition. Mason, OH: Thomson South-Western.

Gilman, R., & Anderman, E. M. (2006). The relationship between relative levels of motivation and intrapersonal, interpersonal, and academic functioning among older adolescents. *Journal of School Psychology, 44,* 375-391.

Goff, M., & Ackerman, P.L. (1992). Personality-intelligence relations: Assessing typical intellectual engagement. *Journal of Educational Psychology, 84*, 537-552.

Hogan, R. T. (1996). A socioanalytic perspective on the five-factor model. In J. S. Wiggins (Ed.), The five-factor model of personality: Theoretical perspectives (pp. 163-179). New York: Guilford Press.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86*, 897-913.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology,* 67, 219-229.

Judge, T. A., Bono, J. E., & Locke, E. A. (2000). Personality and job satisfaction: The mediating role of job characteristics. *Journal of Applied Psychology, 85,* 237-249.

Judge, T.A., Colbert, A.E., & Ilies, R. (2004). Intelligence and leadership: A quantitative review and test of theoretical propositions. *Journal of Applied Psychology, 89*, 542-552.

Judge, T. A. & Ferris, G. R. (1993). Social context of performance evaluation decisions. *Academy of Management Journal*, 35, 80-105.

Judge, T.A., Piccolo, R.F., & Ilies, R. (2004). The forgotten ones? The validity of consideration and initiating structure in leadership research. *Journal of Applied Psychology, 89*, 36-51.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64*, 311-337.

Kristof-Brown, A. L., & Stevens, C. K. (2001). Goal congruence in project teams: Does the fit between members' personal mastery and performance goals matter? *Journal of Applied Psychology, 86*, 1083-1095.

Lounsbury, J. W., Huffstetler, B. C., Leong, F. T., & Gibson, L. W. (2005). Sense of identity and collegiate academic achievement. *Journal of College Student Development, 46,* 501–514.

McCarthy, J.M., & Goffin, R.D. (2001). Improving the validity of letters of recommendation: An investigation of three standardized reference forms. *Military Psychology, 13*, 199-222.

McCrae, R. R., & Costa, P. T. (1989). The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of Personality and Social Psychology, 56,* 586-595.

McCrae, R.R., & Weiss, A. (2007). Observer ratings of personality. In R.W. Robins, R.C. Fraley, & R.F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 259-272). New York, NY, US: Guilford Press.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599-616.

Mohammed, S., Mathieu, J. E., & Bartlett, A. L. (2002). Technical-administrative task performance, leadership task performance, and contextual performance: Considering the influence of team- and task-related composition variables. *Journal of Organizational Behavior, 23*, 795-814.

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

Morgeson, F.P., Reider, M.H., & Campion, M.A. (2005). Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. *Personnel Psychology, 58*, 583-611.

Mount, M. K., Barrick, M. R., & Strauss, J. P. (1994). Validity of observer ratings of the big five personality factors. *Journal of Applied Psychology, 79*, 272-280.

Norman, W.T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology, 66*, 574-583.

Ones, D.S. (1993). *The construct validity of integrity tests*. Unpublished doctoral dissertation.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995-1027.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598-609.

Paunonen, S.V., & Ashton, M.C. (2001). Big Five predictors of academic achievement. *Journal of Research in Personality, 35*, 78-90.

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

Salgado, J.F. (1997). The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology, 82*, 30-43.

Schmidt, F. L, & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmidt, F.L., & Zimmerman, R.D. (2004). A counterintuitive hypothesis about employment interview validity and some supporting evidence. *Journal of Applied Psychology, 89,* 553-561.

Small, E. E., & Diefendorff, J. M. (2006). The impact of contextual self-ratings and observer ratings of personality on the personality-performance relationship. *Journal of Applied Social Psychology, 36*, 297-320.

Society for Human Resource Management. (2005). *Reference and background checking survey report*.

Taylor, P.J., Pajo, K., Cheung, G.W., & Stringfield, P. (2004). Dimensionality and validity of a structured telephone reference check procedure. *Personnel Psychology, 57,* 745-772.

Tupes, E.C., & Christal, R.E. (1961). *Recurrent personality factors based on trait ratings*. (Technical Report No. ASD-TR 61-97). Lackland Air Force Base, TX: U.S. Air Force.

Viswesvaran, C., Ones, D. S, & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557-574.

Wernimont, P.F., & Campbell, J.P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372-376.

Zimmerman, R.D. (2008). Understanding the impact of personality traits on individuals' turnover decisions: A meta-analytic path model. *Personnel Psychology, 61*, 309-348.

Zimmerman, R.D., Mount, M.K., & Goff, M. III. (2008). Multisource feedback ratings and leaders' goal performance: Moderating effects of rating purpose, rater perspective, and performance dimension. *International Journal of Selection and Assessment, 16*, 121-134.

Table 1

Means, Standard Deviations, Inter-correlations, and Reliabilities.

Variables	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. Composite of Four Predictors	4.28	.26	.80									
2. Observer-Rated Conscientiousness	4.41	.27	.80*	.71								
3. Observer-Rated Emotional Stability	4.18	.28	.71*	.51*	.68							
4. Observer-Rated Leadership	4.46	.33	.81*	.59*	.33*	.87						
5. Observer-Rated Interpersonal Skills	4.08	.40	.85*	.53*	.44*	.66*	.76					
6. General Mental Ability	637	48	.09	.08	.01	.12	.08	--				
7. Familiarity with Ratee	4.36	.39	.47*	.38*	.22*	.52*	.39*	.18*	--			
8. Academic Performance	3.46	.38	.19*	.16*	.02	.19*	.22*	.23*	.19*	--		
9. Team Performance	4.47	.41	.22*	.27*	.30*	.05	.11	.05	-.02	.16*	.95	
10. Work Performance	4.05	.37	.42*	.35*	.26*	.51*	.29*	-.12	.18	-.18	.01	.81

Note: *N* = 127 except for Work Performance (*N* = 44). *: *p* ≤ .05. Coefficients alpha are on the diagonal.

Table 2

Results of Regression Analyses Regressing Academic Performance, Team Performance, and Work Performance on Observer-Rated Predictors

Model 1

Step		Dependent Variable		
	Independent Variable	Academic Performance	Team Performance	Work Performance
1	Conscientiousness	.10	.26*	-.01
	Emotional Stability	-.13	.23*	-.10
	Leadership	.06	-.17	.58*
	Interpersonal Skills	.19	-.02	-.16
	R^2 (R)	.07* (.25*)	.13* (.36*)	.27* (.52*)

Model 2

Step		Dependent Variable		
	Independent Variable	Academic Performance	Team Performance	Work Performance
1	General Mental Ability	.23*	.05	-.12
2	Conscientiousness	.09	.26*	-.05
	Emotional Stability	-.12	.23*	.10
	Leadership	.04	-.18	.58*
	Interpersonal Skills	.18	-.02	-.12
	R^2 (R)	.11* (.32*)	.13* (.36*)	.29* (.54*)
	ΔR^2 from Step 1 to 2	.05	.13*	.27*

Model 3

Step		Dependent Variable		
	Independent Variable	Academic Performance	Team Performance	Work Performance
1	General Mental Ability	.23*	.05	-.12
2	Familiarity with Ratee	.16*	-.03	.22
3	Conscientiousness	.08	.27*	-.05
	Emotional Stability	-.12	.23*	.11
	Leadership	.00	-.13	.59*
	Interpersonal Skills	.18	-.02	-.12
	R^2 (R)	.11* (.33*)	.14* (.37*)	.29* (.54*)
	ΔR^2 from Step 2 to 3	.03	.14*	.22*

Note: Standardized beta weights are reported.
N = 127 except for Work Performance (N = 44).
*: $p \leq .05$