7-9-2010

The Role of Rater Motivation in Personnel Selection Validation Studies

Dan Ispas
University of South Florida

The Role of Rater Motivation in Personnel Selection Validation Studies

by

Dan Ispas

A dissertation submitted in partial fulfillment
of the requirements for the degree
Doctor of Philosophy
College of Arts and Sciences
University of South Florida

Co-Major Professor: Walter C. Borman, Ph.D.
Co-Major Professor: Russell E. Johnson
Jennifer K. Bosson, Ph.D.
Edward L. Levine, Ph.D.
Joseph A. Vandello, Ph.D.

Date of Approval
July 9, 2010

Keywords: field study, intervention, Romania, assessment, job performance

Table of Contents

List of Tables

List of Figures

The Role of Rater Motivation in Personnel Selection Validation Studies

Dan Ispas

Abstract

Personnel selection validation studies are routinely conducted in contemporary organizations for selecting and placing employees. Although numerous studies have been conducted with the goal of identifying new predictors, less research was focused on the criterion side. In the current paper, across three studies and five samples, I examined the role played by rater motivation in validation studies. I proposed that rater motivation would impact criterion-related validity of various predictors, the reliability, and the variance of performance ratings. In Study 1, these hypotheses were tested in two samples with varied operationalizations of predictors and of rater motivation. In Study 2, I developed and tested a theoretically based brief intervention designed to increase rater motivation. Study 3 examined directly the link between rater motivation and accuracy.

The results suggest that rater motivation is important and should be considered in validation studies. Rater motivation impacted the criterion related validity of the predictors and the reliability of the ratings. Also, motivated raters showed higher convergence between subjective and objective ratings. The intervention resulted in increased response rates and more reliable ratings. Strengths, limitations and directions for future research are discussed.

Introduction

Performance management systems and performance appraisals are widely used in organizations to inform personnel decisions about compensation, promotions, and employee training and development (Cleveland, Murphy, & Williams, 1989). Much of the early research on performance appraisals focused on psychometric properties of various rating formats, followed later by studies of the accuracy of performance evaluations (e.g., Borman, 1975, 1977). As a result of an influential review published by Landy and Farr (1980), the focus shifted to the cognitive processes involved in observing and evaluating job performance (e.g., DeNisi, 1996). In the early 1990s, several reviews (Bretz, Milkovich, & Read, 1992; Ilgen, Barnes-Farrell, & McKellin, 1993; Murphy & Cleveland, 1991) called for more research on the organizational context in which performance appraisals are conducted. One topic of research that emerged following these calls was rater motivation (Levy & Williams, 2004). Most of the research conducted under the cognitive paradigm assumed that the raters were motivated to give accurate ratings and that the problems in appraisals were caused by cognitive processing errors and complexities (Levy & Williams, 2004). While rater motivation has been examined previously, existent research on rater motivation has mostly ignored two issues: (i) the impact of rater motivation on validity coefficients in validation studies, and (ii) the efficacy of field interventions designed to increase rater motivation.

The goal of this research is to examine the impact of rater motivation on criterion related validity. I propose that rater motivation will moderate the criterion-related validity of various predictors. In Study 1, this hypothesis will be tested in two samples with varied operationalizations of predictors and of rater motivation. In essence, the two samples represent constructive replications of one another (Lykken, 1968). In Study 2, I will develop and test a theoretically based intervention designed to increase rater motivation. Examining rater motivation and developing an intervention to increase such motivation is important for both theoretical and applied reasons. Study 3 will examine directly the link between rater motivation and accuracy (defined as convergence between subjective and objective measures of job performance). Most of the work on validation studies seems largely atheoretical. However, in the current paper I draw from dual-processing theory and leverage-salience theory to explain how rater motivation impacts validation studies and to test an intervention designed to increase rater motivation. From an applied perspective, validation studies have very important consequences for organizations and their members in terms of the selection and placement of new employees. Results of validation studies are used to determine the cut-off scores used for selection measures, to decide who gets a job offer or not, to determine the type of position a person gets hired into (e.g., managerial vs. non-managerial). Traditionally, research on personnel selection tended to be more focused on developing and improving predictors (e.g., increasing validity coefficients, reducing adverse impact, reducing faking on non-cognitive predictors) to the exclusion of criteria. However, if the criteria used in validation research are of low quality, the efforts aimed at improving the predictors only take care of half of the equation. The current paper addresses this problem directly by

examining the criterion in validation studies and by developing a theoretically-based intervention aimed at improving the criterion.

Validation Studies: History and Importance

Substantial evidence has been accumulated in the past century on the criterion-related validities of various predictors such as cognitive ability, personality traits, biodata, interviews, integrity tests, assessment centers, and situational judgment tests (Barrick & Mount, 1991; Barrick, Mount & Judge, 2001; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007). Both professional (SIOP Principles, 2003) and legal guidelines (Uniform Guidelines, 1978) recommend conducting validation studies before including new predictors in the selection process. Despite the number of validation studies, very little research has focused on the validation studies themselves. Russell et al. (1994) conducted a meta-analysis of validation studies published in *Journal of Applied Psychology* and *Personnel Psychology* between 1964 and 1992. They found that several investigator characteristics moderated criterion-related validities, such as impetus behind research (higher validities were obtained for studies addressing an organizational need compared to studies conducted only for research purposes), investigator interests (studies addressing EEO concerns or focused on augmenting a selection system had higher validities than those concerned with maximizing validities and those testing theories), and authors' place of employment (industry authors obtained higher validities compared to academic authors). In addition, Maier (1988) found that the conditions under which the validation study is conducted can impact the validity coefficient. Validity coefficients increased from .09 to .49 and from

.17 to .37 in his two samples by introducing quality controls (e.g., standardizing test administration conditions) in the measurement of the criterion.

The Ratee Perspective

The role of motivation in validation studies has been previously investigated from the perspective of ratees. This stream of research focused mostly on the differences between concurrent and predictive validation strategies. In concurrent designs, also called "present employee method" (Gatewood & Feild, 2005), data on the predictor and the criterion is obtained from a current group of incumbent employees usually at around the same time. One of the criticisms raised against concurrent validation designs is focused on the motivation of the ratees. Since they are already employed, it is conceivable that the ratees are less motivated than job applicants to "do their best" when answering items on predictor tests. Participants in predictive designs, who are actual job applicants in a high-stakes situation, are likely to be more motivated to do well on the tests compared to job incumbents. While intuitive, the prediction that validity differences exist between concurrent and predictive designs has received limited empirical support. For example, Schmitt, Gooding, Noe, and Kirsch (1984), in a meta-analytic review of 99 validation studies published between 1964 and 1982 in *Journal of Applied Psychology* and *Personnel Psychology*, found that studies employing concurrent strategies had a meta-analytic validity coefficient of .34 compared to a validity coefficient of .30 for studies using predictive strategies.

However, missing in these studies of concurrent versus predictive designs were direct measures of ratee motivation. Arvey, Strickland, Drauden, and Martin (1990) therefore specified the construct of test-taking motivation and developed a

multidimensional scale for its measurement, the Test Attitude Survey (TAS). They found that applicants reported greater levels of motivation compared to incumbents. They also examined the moderating effect of TAS in the relationship between predictor scores (ability tests) and supervisor-rated job performance, yet concluded that "the data resulting from the investigation of the TAS factor scores as potential moderators also showed very limited evidence substantiating their use as moderators." (Arvey et al., 1990, pp. 710-711). Since the sample size they used for their analysis was small (N = 69), they encouraged replications with larger samples. Answering this call for more research, Schmit and Ryan (1992) collected data from a sample of 157 undergraduates in a simulated, multi-organization employment system. They found that scores on TAS moderated relationships between predictors and the performance criterion (GPA), but the moderating effect differed depending on the type of predictor. In the case of personality inventories, the validity coefficient was higher for ratees with lower scores on test-taking motivation, whereas the opposite effect was found for a total ability test (the sum of the verbal and quantitative scores of the School and College Ability Test, Educational Testing Service, 1973). Thus it does appear that ratee motivation impacts relationships between predictor and criterion scores.

A Model of Rater Motivation

Examining ratees and their levels of motivation when providing predictor scores is, however, only half of the equation. Attention must also be paid to the motivation of raters who are responsible for providing criterion scores. For the purposes of the current paper, by rater motivation I refer to motivation to engage in the rating process and to provide accurate ratings. Having motivated ratees and valid measures of predictor

variables does little good if criterion scores are flawed. Unfortunately, several problems with criterion variables and measures have been identified (Austin & Villanova, 1992). For example, a common complaint of performance appraisal systems is the rater's tendency to give high ratings. It's not unusual for a large majority of employees to receive extremely high ratings (e.g., 80% of employees receiving a rating a 6 or 7 for ratings done on a 7 point scale – Murphy and Cleveland, 1995). In fact, from the rater's perspective, there are very few reasons in favor of giving accurate ratings (Murphy & Cleveland, 1995). As discussed below, there are more reasons in favor of giving inaccurate ratings (there are limited rewards and more negative consequences for accurate ratings). As such, it appears that rater leniency is not error but a conscious effort on the part of the rater. Evidence for existence of rater leniency has been found in both primary studies and meta-analyses. For example, Harris, Smith, and Champagne (1995) found that ratings made for administrative purposes were higher that the ratings made for research purposes. In a meta-analytic study, Jawahar and Williams (1997) examined this "performance appraisal purpose effect" using data from 22 studies with a total sample size of 57,775. They found that ratings made for administrative purposes are approximately one third of a standard deviation higher than those made for research purposes.

The first performance appraisal model that explicitly included rater motivation was the one proposed by DeCotiis and Petit (1978). They proposed six determinants of rater motivation at the rating stage: perceived consequences of accurate appraisals for both the rater and the ratee, rater perception of the adequacy of the instrument used in appraisals, organizational policies and practices, the rating format, availability of

standards of performance, and purpose of the appraisal. More recently, Harris (1995) developed a theoretical model of the determinants of rater motivation that goes beyond just the rating stage. The performance appraisal process consists of a series of steps: observing the behavior of ratees, storing, retrieving, and integrating information regarding the ratee, providing a rating, and delivering feedback (Wexley & Klimoski, 1984). Harris (1995) proposed three determinants of rater motivation: perceived rewards, perceived negative consequences, and impression management concerns. These determinants are affected by situational (e.g., accountability, organizational HRM strategy, trust) and personal factors (e.g., self-efficacy, mood). The three determinants are discussed next.

Rewards are an important determinant for almost any behavior (Kanfer, 1990). Surprisingly, in the performance appraisal context, rewards for providing accurate ratings are rarely used despite the fact that these ratings are used to make important organizational decisions (e.g., pay raises and promotions, termination and downsizing decisions). Providing accurate ratings can be rewarded via extrinsic and intrinsic rewards. *Extrinsic rewards* refer to the attainment of valuable outcomes (bonuses, pay raises, promotions) in exchange for providing accurate ratings. Field research on the extrinsic rewards is limited and portrays a "dismal picture" (Harris, 1995, p. 740). A paper by Napier and Latham (1986) examined the rater's expected outcomes for providing feedback to employees across two studies. Most rater responded that the primary result would be "nothing." *Intrinsic rewards* refer to the fact that some raters may find engaging in performance appraisal activities as an activity inherently satisfying. For example, research by Rand and Wexley (1975) suggests that raters tend to view their

subordinates more similar to themselves than they actually are so raters that see similarities between themselves and their subordinates will tend to give favorable ratings. Also, both raters and rates dislike giving and receiving negative feedback (Fisher, 1974).

Negative consequences are the second determinant of rater motivation proposed by Harris (1995). Negative consequences can be organized into five categories: damage to the relationship between rater and ratee, negative impact on employees' morale, criticism from the subordinate, criticism from the rater's supervisor, and interference with other tasks. By giving accurate (which often means deflated or lower-than-expected) ratings, the raters are concerned that they may negatively impact the relationship they have with the ratees and that they may even demoralize the ratees (Longenecker et al., 1987; Murphy & Cleveland, 1991). Criticism from the ratees is also a possibility and it can result in legal action against the rater and/or the organization. Also, the jobs of most managers are comprised of several responsibilities, many of which are perceived to be more important than rating employees. Thus, managers may therefore decide to allocate little time and effort to the task of rating subordinate performance (Bernardin & Villanova, 1986).

The third determinant of rater motivation proposed by Harris (1995) is the raters' concern for impression management. Zerber and Paulhus (1987) distinguished between two types of impression management: self-impression management and management of others' impressions. Self-impression management is related to rater motivation in three ways: (i) by rating their subordinates highly, managers may perceive themselves as successful managers, (ii) in order to maintain a positive view of the organization, raters may also believe that the mere fact that the ratee works there is evidence that the ratee

has satisfactory job performance, and (iii) raters may view their role as manager differently in that not all raters believe that it is their role to provide accurate ratings. In terms of others' impression of oneself, raters may feel that by giving low ratings, their competence will be questioned by their own supervisor(s). Also related to others' impression of oneself are organizational norms. For example, some organizations may have norms that encourage high, inaccurate ratings, and thus managers who rate their subordinates in this way are behaving according to organizational practices.

Although Harris' (1995) determinants were discussed in reference to the general purpose of rating for administrative purposes in organizations, we can apply these determinants to the case of validation studies. A typical validation study involves several steps: a job analysis, identification of relevant performance dimensions, identification of the knowledge, skills, and abilities (KSAs) needed for the job, development of assessment devices for the measurement of KSAs, evaluation, and implementation of the new system (Gatewood & Feild, 2005). Data is collected from job incumbents, who are administered the predictor measures, and their supervisors, who provide performance ratings. Participation by both the raters and the rates is usually voluntary. The validation study is usually conducted by an outside consultant who is working with the top management of the organization and the raters are rarely involved in the decision making part of the process. It is usually difficult for both the incumbents and the raters to see the importance of the validation study. There are usually no immediate or direct rewards for the raters, because most validations studies are perceived as research projects. Raters have to take the time from their busy schedules to rate subordinates—usually multiple subordinates—on an ostensibly pro bono basis. There are no negative consequences for

raters since participation is voluntary which means zero or very limited accountability. Also, there are some impression management concerns since participation in the validation study is usually requested by a high-level person in the organization (e.g., Chief HR Officer). Management of others' impressions and self-impression management can be alleviated by responding to the request to participate. However, just responding to the request to participate does not mean that the raters will expend the resources necessary for accurate ratings, especially since managers usually have to rate multiple subordinates. Taken together, then, raters likely have minimal motivation to provide accurate ratings for validation studies. While most of the research has focused on the differences in ratings when ratings are made for administrative versus research purposes, I am not aware of any research that has examined the impact of rater motivation on the criterion related validity of various predictors used.

Rater Motivation in Validation Studies

So far, research on rater motivation has focused on the mean differences between ratings for developmental or research purposes and ratings for administrative purposes. As reviewed above, a consistent finding of this line of inquiry is that raters are more lenient (i.e., give higher mean ratings) when the ratings are used for administrative purposes. Research on rater motivation appears to suffer from the same problem identified by Arvey et al. (1990) for research focused on ratee motivation. That is, rater motivation is not measured directly and it is merely assumed that raters providing ratings for research purposes are more motivated to give accurate ratings than when giving ratings for administrative purposes. The focus of this paper is on the role rater motivation plays in validation studies, where ratings are used for research purposes. I propose that

rater motivation will moderate relationships between predictors and criteria, such that these relationships will be stronger when rater motivation is high versus low.

Motivated raters devote more attentional and cognitive resources to the task at hand—that is, they engage in more symbolic or explicit cognitive processing. Explicit processing is characterized by being conscious, relatively slow, and effortful (e.g., Lieberman, 2007; Satpute & Lieberman, 2006; Stanovich, 1999, 2004). People engaged in explicit processing are less likely to fall prey to biased decision-making and memory heuristics. Human thinking often relies on the operation of intuitive heuristics instead of deliberate and controlled reasoning because humans have finite amounts of attentional and cognitive resources that can be allocated to decision-making. When faced with making hundreds or thousands of decisions over the course of a day, humans lack the computational power to bring explicit processing to bear on all of these decisions. Thus, much of the work is carried out by heuristic or implicit information processing, which occurs quickly and with little effort. However, one consequence of heuristic-based processing is that it may generate answers that are logically or probabilistically incorrect (e.g., Evans, 2002; Kahneman, Slovic, & Tversky, 1982). This line of reasoning is consistent with work on dual-processing theories (Strack & Deutsch, 2004). According to dual-processing theories (e.g., Johnson et al., 2010; Strack & Deutsch, 2004), there are two modes of cognitive processing: implicit and explicit. In order to operate, explicit processing requires sufficient attention, motivation, and capacity (i.e., necessary time and resources). It refers to a slower conscious process where information is from working memory. In contrast, implicit processing requires few resources and often occurs automatically.

Outcomes of explicit processing are usually superior to those of implicit processing: alternative solutions to problem-solving and reasoning, better organization of information and integration in memory, a greater likelihood of attitude and behavior change, less use of stereotypes in judgments, and facilitated learning of new facts and rules (e.g., Anseel, Lievens, & Schollaert, 2009; Smith & DeCoster, 2009). Attitudes formed as a result of using explicit processing are more predictive of behavioral intentions and actions, and are more persistent over time (for reviews, see Cacioppo, Petty, Feinstein, Blair, & Jarvis, 1996; Petty, Wegener, & Fabrigar, 1997). When raters have low motivation, the performance ratings they give to their subordinates are more likely to reflect implicit performance theories and decision-making heuristics rather than factual summaries of actual performance. For example, the ratings given when motivation is low may be adversely impacted by primacy and recency heuristics. That is, unmotivated raters may be more likely to remember a person's initial and most recent performance behaviors, but fail to recall behaviors that occurred in between. Unmotivated raters could also choose salient samples of performance (either very good or very bad) and use those as a basis for their ratings. When engaging in rating, raters are theorized to form schemas to categorize their subordinates. The term "cognitive miser" refers to a widely cited schema function from social psychology. To reduce the overall processing load, people use ''shortcuts'' to conserve mental resources when they are trying to make sense of other people (Fiske & Taylor, 1984). When making performance appraisals, raters have a considerable cognitive load: they usually have to consider multiple subordinates and multiple situations. Motivated raters are more likely to use reflective/explicit processing as opposed to unmotivated raters who are more likely to use

impulsive/implicit processing. Motivated raters will expand the resources needed to give accurate ratings, as such there will be a strong association between the employees scores on the predictor and their performance ratings. On the other hand, for unmotivated raters the correlation between predictor scores and performance ratings will be reduced due to inaccurate ratings. Unmotivated raters may, therefore, simply take the ‘‘easy’’ way out, by giving all their subordinates an average rating or using another, similar tactic, to avoid investing effort and time into rating (Harris, Ispas, & Schmidt, 2008).

Therefore, the following hypothesis is proposed:

Hypothesis 1: Rater motivation will moderate relationships between predictors and criteria such that these relationships will be stronger when rater motivation is high versus low.

Also, raters who are motivated will discriminate more among the performance of their subordinates when provided ratings compared to raters who are unmotivated. If so, then there should be greater variance in ratings provided by motivated versus unmotivated raters. As such, I also hypothesize the following:

Hypothesis 2: There will be greater variance in performance ratings when raters have high versus low motivation.

An important criterion for the evaluation of job performance is the reliability of the ratings (Viswesvaran, 2001). Reliability refers to consistency of measurement (Nunnally, 1978). Motivated raters should be more consistent in their ratings than unmotivated raters. Therefore, I hypothesize the following:

Hypothesis 3: The reliability of the ratings made by motivated raters is higher than the reliability of the ratings made by raters low in motivation.

Plan of the current research

The first two studies will follow a constructive replication format (Lykken, 1968), meaning that I will vary my conceptualization and operationalization of the predictor and moderator variables. In Study 1, sample 1 the predictor will be a cognitive ability test, while rater motivation and accountability will be measured using a one-item self-report scale. In Study 1, sample 2 the predictor will be a job knowledge test, while rater motivation will be measured unobtrusively and more objectively as the time spent making the ratings. Study 2 will present a field experiment of an intervention designed to increase rater motivation. Study 3 will examine if rater motivation increases the rater's accuracy by examining if motivated raters have a higher association between subjective and objective measures of job performance.

Study 1 Method

The data for both samples used in Study 1 were collected as part of validation studies conducted for the purpose of identifying new predictors to be used for personnel selection.

Sample 1 Participants and Procedure

The participants were 220 employees and their supervisors recruited from a Romanian manufacturing organization. The majority of employees were male (61%) and their age ranged from 22 to 55 years. The data on the predictor was collected in small groups of 15-20 participants using paper and pencil questionnaires.

Sample 1 Measures

Predictor. The predictor is the General Ability Measure for Adults (GAMA; Naglieri & Bardos, 1997), a non-verbal cognitive ability test that consists of 66 items grouped in four subtests: Matching (11 items), Analogies (17 items), Sequences (20 items), and Construction (18 items). For the Matching items, the respondents examine the shape, color and configuration of a stimulus item to determine the correct response option. For the Analogies items, the respondents must identify the pattern of the relationship between a pair of abstract figures and recognize a similar relationship in a different pair of figures. For the Sequences items, the respondents must identify the pattern of change in a configuration of figures as they move through space. The Construction items require analyzing, synthesizing, rotating, and combining a number of

shapes to mentally construct a figure identical to one of the response options. The GAMA is unidimensional, has a split-half reliability around .90, and test-retest reliabilities ranging between .67 and .84. Iliescu and Ghinta (2008) and Naglieri and Bardos (1997) reported supportive evidence for the convergent validity of this instrument because scores on the GAMA were significantly correlated with scores on other cognitive ability tests, including the Wonderlic Personnel Test (Wonderlic Inc, 2002), Shipley Institute of Living Scale (Shipley, 1991), Kaufman Brief Intelligence Test (Kaufman & Kaufman, 1990), and Multidimensional Aptitude Battery (Jackson, 2003; *r*s were around .70). In terms of criterion-related validity, scores on the GAMA have been found to predict academic achievement (Bardos, 2003; Crawford et al., 1999) and job performance (Ispas, Iliescu, Ilie, & Johnson, 2010).

Criterion. Supervisors rated the job performance of their subordinates using a 6-item behaviorally-anchored rating scale. The six items cover the major dimensions of performance such as problem solving, effort, interactions with co-workers, and overall job performance. Supervisors rated employees' performance on a 1–7 scale, where high (6–7), moderate (3–5), and low (1–2) performance levels were delineated (see Judge & Erez, 2007). The items were selected by the participating organization.

Moderator. Supervisors responded to a single item adapted and modified from the Motivation subscale of the Test Attitude Survey (Arvey et al., 1990). The item was: "I tried to do the very best I could on these ratings." A 5-point Likert scale was used ranging from 1 = strongly disagree to 5 = strongly agree.

Sample 2 Participants and Procedure

The participants were 300 incumbents and their supervisors recruited from a Romanian manufacturing organization (different than sample 1). The majority of participants were male (89%) and their age ranged from 18 to 59 years. The data on the predictor was collected in small groups of 15-20 participants using paper and pencil questionnaires.

Sample 2 Measures

Predictor. A 20-item job knowledge test was developed following the procedures outlined by Muchinsky (2004). Working with subject matter experts (SMEs), 35 items were written to cover the content domain and pilot tested using a small validation sample (N = 45). Based on an item analysis and the organization's request the number of items was reduced to twenty.

Criterion. Supervisors rated the job performance of their subordinates using a 4-item behaviorally-anchored rating scale. Similar to Sample 1 the four items cover the major dimensions of performance such as problem solving, effort, and overall job performance. The items were selected by the participating organization.

Moderator. The ratings of job performance made by supervisors were collected on-line. The amount of time (in seconds) spent on-line by the rater when filling out the survey was used as an objective and unobtrusive measure of rater motivation. The assumption is that motivated raters will spend more time reading and responding to the survey items than unmotivated rates. I tested this assumption in a pilot study using 41 raters from a Romanian organization (different than the organizations used in Samples 1 and 2). The participants were asked to rate one of their subordinates using an on-line

survey and the job performance measure from Sample 2. Rater motivation was measured using both the direct method from Sample 1 and the unobtrusive measure from Sample 2. The direct method consisted of a single item adapted and modified from the Motivation subscale of the Test Attitude Survey (Arvey et al., 1990). The item was: "I tried to do the very best I could on these ratings." A 9-point Likert scale was used ranging from 1 = strongly disagree to 9 = strongly agree.) The unobtrusive measure was the time spent on-line by the rater. The two measures of rater motivation were correlated at $r = .65$, $p < .001$, providing evidence for the construct validity of the unobtrusive measure of rater motivation. A similar measurement of motivation was recently used by Oppenheimer, Meyvis, and Davidenko (2009).

Study 1 Results

Means, standard deviations, reliabilities and inter-correlations are presented in Table 1 for Sample 1 and in Table 2 for Sample 2. Although not the focus of the current study, both predictors (cognitive ability and job knowledge) were significantly related to job performance at levels similar to previous studies (see Schmidt & Hunter, 1998). Specifically, cognitive ability had a correlation of .41 ($p < .01$) with job performance, while job knowledge was correlated with job performance at .35 ($p < .01$).

Table 1

Descriptive and Correlational Information for Study 1 Sample 1

	M	SD	1	2	3
1. Predictor: Cognitive Ability	37.68	11.49	*(.86)*		
2. Criterion: Job Performance	18.24	4.16	.41**	*(.72)*	
3. Rater motivation	2.95	1.17	-.03	.25**	*(NA)*

** $p < .01$, $N = 220$

Table 2

Descriptive and Correlational Information for Study 1 Sample 2

	M	SD	1	2	3
1. Predictor: Job Knowledge Test	11.97	4.88	*(.90)*		
2. Criterion: Job Performance	12.73	3.82	.34**	*(.82)*	
3. Rater motivation	255.56	132.53	-.02	.02	*(NA)*

** $p < .01$, $N = 300$

Hypothesis 1 proposed that rater motivation will moderate relationships between predictors and criteria such that these relationships will be stronger when rater motivation is high versus low. This hypothesis was tested using hierarchical moderated multiple regressions (Cohen, Cohen, West, & Aiken, 2003). The interaction term was created by multiplying the predictor variable (cognitive ability in Sample 1 and job knowledge in Sample 2) and the moderating variable (rater motivation). I then conducted a hierarchical regression analysis by regressing job performance on the predictor variable in Step 1, rater motivation in Step 2, and the interaction term in Step 3. For Sample 1, the interaction term explained a significant amount of variance: F (1, 216) = 11.92, $p < .001$, ΔR^2=0.04. The interaction was examined further by conducting simple slope tests (Cohen et al., 2003). The results show that the relationship between the predictor (cognitive ability) and job performance is stronger when rater motivation is high ($\beta = .60$, $p < .001$) compared to when rater motivation is low ($\beta = .22$, $p < .01$). The full results of the regression analyses for Sample 1are presented in Table 3. A plot of the interaction can be seen in Figure 1.

Table 3

Rater Motivation as Moderator in Sample 1 Study 1

Step	Independent variable	*B*	*SE B*	95% CI	*β*	ΔR^2
1	Predictor	.15	.02	.11-.19	.41**	.17**
2	Rater motivation	.94	.21	.53-1.36	.27**	.07**
3	Predictor x Rater motivation	.06	.02	.03-.09	.81**	.04**

Note: N = 220, CI = confidence interval. The predictors and the moderating variables were standardized. ** $p < .01$

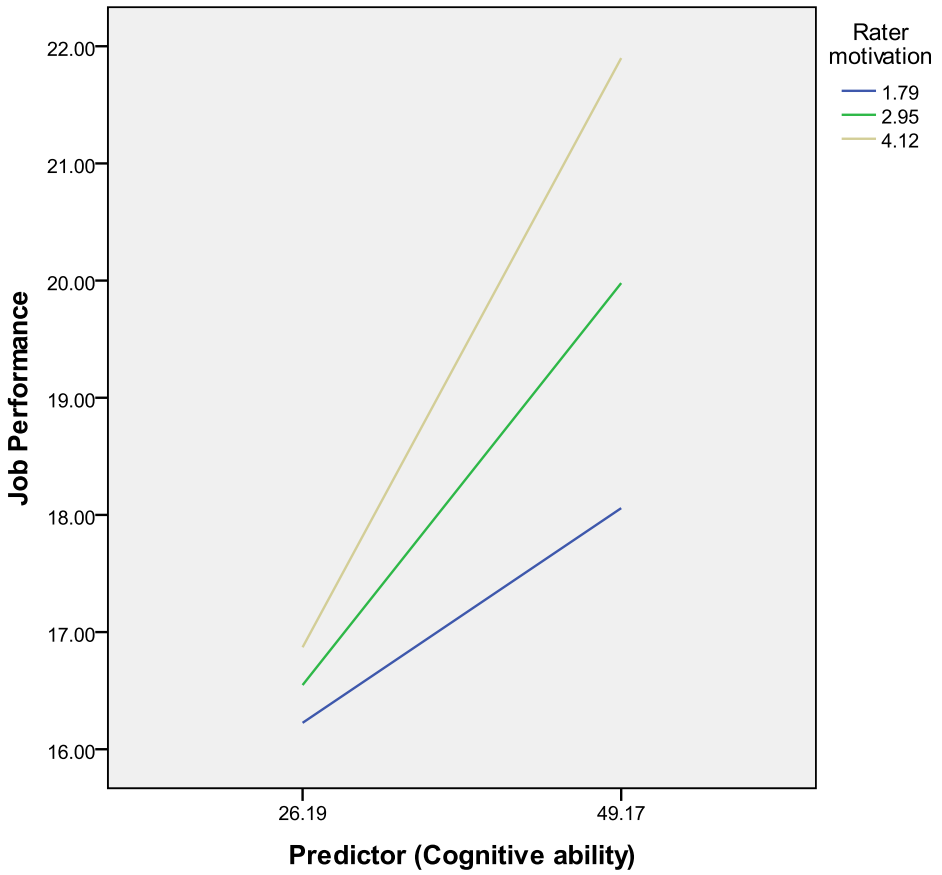

Figure 1: Rater motivation as moderator in Study 1 Sample 1

For Sample 2, similar results were observed as the interaction term explained a significant amount of variance: F (1, 296) = 6.51, $p < .001$, ΔR^2=0.02. Simple slope tests show that the relationship between the predictor (job knowledge) and job performance is stronger when rater motivation is high ($\beta = .47$, $p < .001$) compared to when rater motivation is low ($\beta = .20$, $p < .05$). The full results of the regression analyses for Sample 2 are presented in Table 4. A plot of the interaction is illustrated in Figure 2. Therefore, Hypothesis 1 was supported in both samples.

Table 4

Rater Motivation as Moderator in Sample 2 Study 1

Step	Independent variable	*B*	*SE B*	95% CI	β	ΔR^2
1	Predictor	.27	.04	.18, .35	.34**	.12**
2	Rater motivation	.00	.00	-.00, .00	.03	.00
3	Predictor x Rater motivation	.00	.00	.00, .01	.45*	.02*

Note: N = 300, CI = confidence interval. The predictors and the moderating variables were standardized. * $p < .05$, ** $p < .01$.

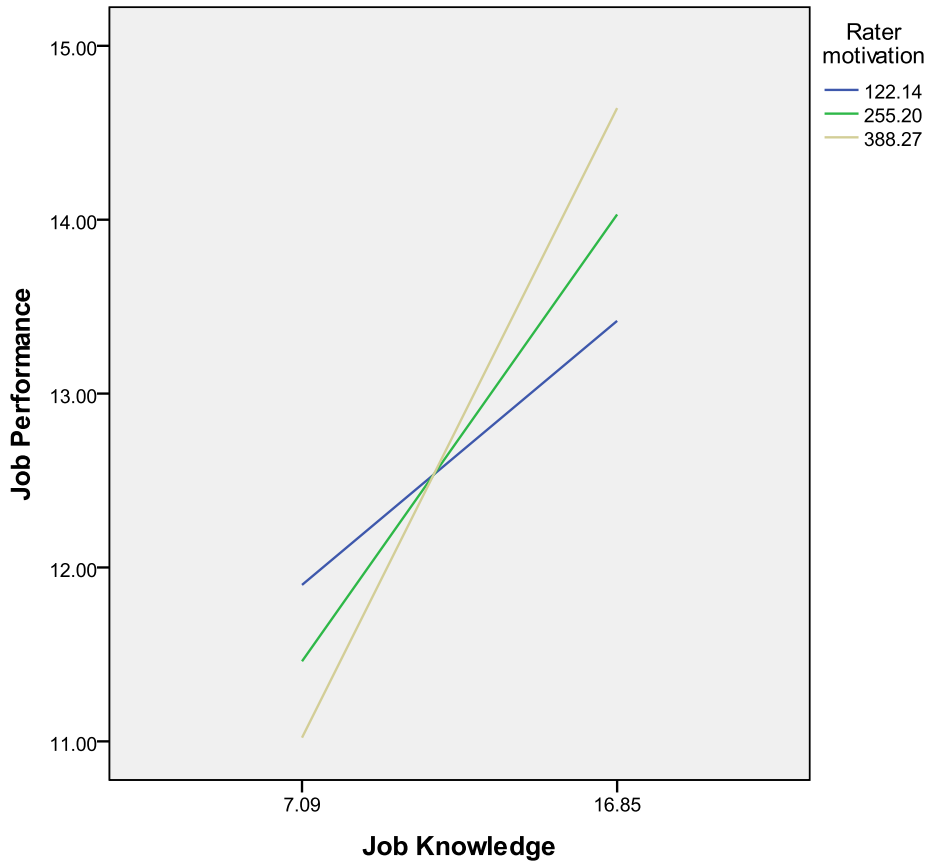

Figure 2: Rater motivation as moderator in Study 1 Sample 2

Hypothesis 2 proposed that there will be greater variance in performance ratings when raters have high versus low motivation. In sample 1, I created two groups, the low motivation group (respondents that answered 1 and 2 on the rater motivation item) and the high motivation group (respondents that answered 4 and 5 on the rater motivation item). For Sample 1, the variance of the low motivation group (N = 86) was 15.90 (M = 17.43, SD = 3.99), while the variance for the high motivation group (N = 64) was 21.65 (M = 19.81, SD = 4.65). Levene's test for the equality of variances was not statistically significant (F = 1.76, p = .187) indicating that the variances were not different. In Sample 2, I created two groups by using a median split. The variance of the low motivation group (N = 151) was 13.06 (M = 12.62, SD = 3.61), while the variance for the high motivation

group (N = 149) was 16.24 (M = 12.85, SD = 4.03). Although the variance in the high motivation group is higher than the variance in the low motivation group, the Levene's test for the equality of variances was not statistically significant (F = 1.03, p = .311). As such, Hypothesis 2 was not supported in neither Sample 1 or in Sample 2.

Hypothesis 3 proposed that the ratings made by high motivation raters will be more reliable (internally consistent) than those made by the low motivation raters. In Sample 1, for the same groups created to test Hypothesis 2, the internal consistency coefficients (Cronbach's alpha) were computed. For the high motivation group, the reliability coefficient was 0.814, whereas the reliability coefficient was 0.675 for the low motivation group. The reliability coefficients were compared using the *k*-samples significance test proposed by Hakstian and Whalen (1976). The method is distributed as a χ^2 distribution with 1 degree of freedom under the null hypothesis of equal reliability. The results show that the difference between the two reliability coefficients was statistically significant: $\chi^2(1) = 5.69$, $p = 0.0171$. In Sample 2, for the same groups created to test Hypothesis 2, the internal consistency coefficients (Cronbach's alpha) were computed. For the high motivation group, the reliability coefficient was 0.856. For the low motivation group, the reliability coefficient was 0.784. The results show that the difference between the two reliability coefficients was statistically significant: $\chi^2(1) = 4.55$, $p = 0.0329$. Therefore, Hypothesis 3 was supported across both samples.

Study 1 Discussion

The purpose of Study 1 was to examine the role played by rater motivation in validation studies. The results suggest that rater motivation acted as a moderator of the validity coefficients such that there was a higher validity coefficient when rater motivation was high. The results were replicated across two samples with different measurements for rater motivation and with different predictors. In both samples, I also found that raters high on motivation had more variance in their job performance ratings; however the differences were not statistically significant. The ratings made by motivated raters were also more internally consistent than those made by raters low on motivation.

Overall, the results of Study 1 highlight the importance of rater motivation in personnel selection validation studies. Motivated raters show more reliable ratings and can get higher validity coefficients. However, one question that remains answered is whether we can increase rater motivation?

Study 2 Introduction

Rater motivation is a critical factor in order for performance appraisals to be effective (Murphy & Cleveland, 1995; Harris, 1995). The results of Study 1 suggest that rater motivation is also critical in validation studies as well. However, the number of interventions designed to increase rater motivation is very limited. Only a few studies have proposed and investigated possible interventions. For example, Roch (2007) hypothesized that rater teams can increase rater motivation. However, she did not find support for this hypothesis. Salvemini, Reilly, and Smither (1993) proposed the use of incentives as a way to increase rater motivation. They found that providing monetary rewards increased the accuracy of raters. Other motivational interventions that have been proposed include potential discussions with the ratee (Ilgen & Knowlton, 1990) and scrutiny by an expert (Mero & Motowidlo, 1995). These interventions, however, were tested in simulated contexts, usually involving samples of undergraduate students as raters. Also, most of these interventions have limited use in validations studies. Therefore, the purpose of Study 2 is to develop and test a theoretically-based intervention designed to increase rater motivation using a field experiment.

Theoretical Basis for the Intervention

Leverage-salience theory of survey participation (Groves, Singer, & Corning, 2000) is a recent application and refinement of the dual-processing theories of persuasion (e.g., Petty & Cacioppo, 1986). According to the leverage-salience theory of survey

participation, respondents vary in the importance they assign to various aspects of survey requests. Importance derives from various features of the survey, respondent, and the situation. For example, importance may depend on whether respondents find the survey topic interesting versus uninteresting, or whether they have sufficient versus inadequate time to participate. When asked to participate in a survey, one of these features will become salient for the participant in his/her interaction with the survey materials and can impact the participant's amount of effort put into responding to the survey. A consistent finding of research in the communication and persuasion literature is that when an issue is perceived as having high personal relevance to individuals they are more likely to carefully examine the content of the information presented. Information that has high personal relevance is more likely to be processed more in-depth via symbolic or explicit processing. As such, a data collector can tailor his/her approach to each respondent or class of respondents (Groves, Singer, & Groning, 2000). As applied to the current study, salience or personal importance refers to the extent to which raters feel that the validation study is important to them or their organizations. By increasing the personal importance of the validation study, raters are expected to be more motivated and engage in more careful processing when they rate their subordinates. In order to increase the personal importance of the validation study, I will highlight the consequences of a properly conducted validation study for both the raters and their organizations. While I acknowledge that most likely a training intervention would be the most effective approach, it was important to find a short, inexpensive intervention that can be embedded in common validation studies.

Based on the discussion above, the following hypotheses are proposed:

Hypothesis 4: The response rate among the raters will be higher in the intervention conditions versus the control conditions.

Hypothesis 5: Rater motivation will be higher in the intervention conditions versus the control conditions.

Hypothesis 6: The criterion-related validity of the predictor will be higher in the intervention conditions versus the control conditions.

Study 2 Method

Participants and Procedure

Three hundred and sixty managers and their subordinates were invited to participate in a validation study of a job knowledge test. The managers received an e-mail containing a link to a web-survey. The raters were randomly assigned to one of the three conditions: a low salience condition and two high salience conditions (120 participants in each condition). The wording for the three conditions is presented below. The high salience conditions differ from each other on their focus on the benefits for individual and the benefits for the organization.

The general instructions were as follows:

We are conducting a research study to identify new tools that can be used for selecting new employees for (name of organization).

An important part of the process requires your participation. After clicking on the link below, you will be taken to a secured website where you will be asked to provide performance ratings for one randomly selected subordinate. The employee will NOT see the ratings you provide.

We would appreciate your taking the time to complete the following survey. It should take about five minutes of your time. Your participation is completely voluntary and there are no risks associated with your participation. Furthermore, all personal identifying information will be kept separate from your responses on the questionnaires.

The wording for the collective condition was:

Why should you participate?

The results of this study will be used to determine selection tools for recruiting and placing new employees in our organization. By carefully responding to the survey, we can be more confident in the decisions we make by hiring people who will fit in well with the company. As such we will be able to hire better employees and colleagues. The decisions made on the basis of this research study will benefit our entire organization.

The wording for the individual condition was:

Why should you participate?

The results of this study will be used to determine selection tools for recruiting and placing new employees in the organization. By carefully responding to the survey, we can be more confident in the decisions we make by hiring people who will work well with you. As such you will have better employees and colleagues working for you. The decisions made on the basis of this research study will benefit you personally.

Measures

Predictor. Job incumbents filled out a 24-item job knowledge test developed following procedures similar to Study 1 - Sample 2.

Criterion. A four-item behaviorally anchored rating scale was used by supervisors to rate the performance of their subordinates. The items were selected by the participating organization.

Moderator. Rater motivation was measured with a five item scale developed by Hedge and Teachout (2000). Example items are: "I care how accurate my ratings were" and "I made the most accurate ratings I could" Raters responded to these items via a five point Likert-type response (from "1 = Strongly disagree" to "5 = Strongly agree").

Manipulation Checks. A one item measure was used to measure personal importance: "How important is participating in this validation study for you?" A one item measure was used to measure organizational importance: "How important is participating in this validation study for your organization?" A nine-point Likert-type response scale was used ranging from "1 = Not at all important" to "9 = Extremely important."

Control Variables. Given that the data was collected from Romania, country which is considered as a collectivistic culture, I included Individualism and Collectivism as a control variable. Individualism and collectivism were measure using four item scales developed by Triandis and Gelfland (1998). Sample items are "I often do my own things" and "I feel good when I cooperate with others". A nine-point Likert-type response scale will be used ranging from "1 = Strongly disagree" to "9 = Strongly agree".

Study 2 Results

Means, standard deviations, reliabilities and inter-correlations are presented in Tables 5 and 6. Due to their low reliabilities, and to the fact that they had no impact on any of the hypotheses tested in this study, the individualism and collectivism scales were dropped from further analyses.

Table 5

Descriptive Statistics for Study 2

	Control (C)		Individual benefits (I)		Organizational benefits (O)	
Rater motivation	9.00_a	3.29	10.85_b	5.31	11.14_b	5.01
Job performance	13.10	3.06	12.77	3.75	12.78	4.35
Job knowledge	10.66	4.52	11.58	4.04	12.37	4.89
Individualism	28.03	4.84	28.44	4.52	29.73	3.77
Collectivism	28.34	3.76	28.06	3.81	28.54	3.54
Manipulation check Individual	3.36_a	1.16	5.20_b	1.56	3.71_a	1.34
Manipulation check Organizational	3.59_a	1.09	3.79_a	1.29	5.71_b	1.35

Note: $N = 226$, * $p < .05$, ** $p < .01$, *** $p < .001$. Means with different subscripts are significantly different at the .05 level (Tukey's HSD)

Table 6

Correlation Matrix and Reliabilities for Study 2

	1	2	3	4	5
1. Job performance	*.84*				
2. Predictor: Job knowledge	.39**	*.94*			
3. Rater motivation	.01	-.01	*.91*		
4. Individualism	.09	.06	-.03	*.57*	
5. Collectivism	.06	-.01	-.09	.49**	*.53*

** $p < .01$, $N = 226$

First, I tested the two manipulation checks. For the personal importance manipulation check ("How important is participating in this validation study for you?"), as expected, respondents in the individual benefits condition reported higher means (M = 5.20, SD = 1.56) than both the control (M = 3.36, SD = 1.16) and the organizational benefits condition (M = 3.71, SD = 1.34); F (2, 223) = 38.62, $p < .001$. Post-hoc tests (Tukey's HSD) showed that there were no statistically significant difference between the organizational benefits condition and the control condition.

Similar results were found for the organizational importance manipulation check ("How important is participating in this validation study for your organization?"). Respondents in the organizational benefits condition reported higher means (M = 5.71, SD = 1.35) than both the control (M = 3.59, SD = 1.09) and the individual benefits condition (M = 3.79, SD = 1.29); F (2, 223) = 65.04, $p < .001$. Post-hoc tests (Tukey's HSD) showed that there were no statistically significant differences between the individual benefits condition and the control condition.

Before testing the main hypotheses proposed in Study 2, in an effort to increase the generalizability of the results, I re-tested Hypotheses 2 and 3 using the Study 2 sample. Recall that Hypothesis 2 proposed that there will be greater variance in performance ratings when raters have high versus low motivation. I compared the variance in job performance ratings between the individual benefits condition and the control condition and found that although the variance was higher in the individual condition (14.09 vs. 9.39) the Levene's test for equality of variance was only statistically significant at the .10 level: $F = 2.89$, $p = .09$. When I compared the variance in job performance ratings between the organizational benefits condition (Variance = 18.94) and the control condition (Variance = 9.39), the Levene's test for equality of variance was statistically significant: $F = 7.20$, $p < .05$. Thus, there was mixed support for Hypothesis 2.

Hypothesis 3 predicted that the reliability of the ratings made by motivated raters is higher than the reliability of the ratings made by raters low in motivation. The reliability coefficients were compared using the *k*-samples significance test proposed by Hakstian and Whalen (1976). When comparing the control condition (alpha = .703) and the individual benefits condition (alpha = .852), the difference between the two reliability coefficients was statistically significant: $\chi^2(1) = 5.98$, $p = 0.0145$. When comparing the control condition (alpha = .703) and the organizational benefits condition (alpha = .889), the difference between the two reliability coefficients was also statistically significant: $\chi^2(1) = 11.24$, $p = 0.0008$. Therefore, Hypothesis 3 was fully supported.

Moving on with the main hypotheses for Study 2, recall that Hypothesis 4 proposed that the response rate among the raters will be higher in the intervention

conditions versus the control conditions. The response rate for the control group was 50.83% (61 out of 120). For the experimental conditions the response rates were: 71.67% (86 out of 120) for the individual benefits condition and 65.83% (79 out of 120) the organizational benefits condition. The response rates were compared using z-tests for proportions. The results show that the response rates were indeed higher for both the individual benefits condition ($z = 3.18$, $p < .05$) and the organizational benefits conditions ($z = 2.23$, $p < .05$). Therefore, Hypothesis 4 was supported.

Hypothesis 5 proposed that rater motivation will be higher in the experimental conditions compared to the control condition. I tested this hypothesis using ANOVA. The results show that the omnibus F test was statistically significant F (2, 223) = 4.00, $p <$ .05. Post hoc tests (Tukey's HSD) show that in both the individual ($M = 10.85$, $SD =$ 5.31) and the organizational ($M = 11.14$, $SD = 5.01$) benefits conditions rater motivation was higher than in the control condition ($M = 9.00$, $SD =$3.29). Thus, Hypothesis 5 was also supported.

Hypothesis 6 proposed that the validity coefficients in the experimental conditions will be higher that the validity coefficients in the control condition. Validity coefficients (correlations between the predictor – the job knowledge test and the job performance ratings) were calculated for each of the three conditions. Results show that the validity coefficients in both the individual benefits condition ($r = .45$, $p < .01$) and the organizational benefits condition ($r = .42$, $p < .05$) were higher than in the control condition ($r = .32$, $p < .01$). However, the differences were not statistically significant $z =$ -0.89, $p = 0.18$, for the individual benefits condition, and $z = -0.66$, $p = 0.25$, for the organizational benefits condition. Therefore, Hypothesis 6 was not supported.

Since there were no differences between the two experimental conditions, I retested Hypothesis 6 by combining the two experimental conditions. The differences between the combined experimental and control conditions were not statistically significant: $z = -0.76$, $p = 0.22$

Study 2 Discussion

The main purpose of Study 2 was to test an intervention designed to increase rater motivation, response rates and validity coefficients. The results show that in both of the experimental conditions (individual and organizational benefits) rater motivation and the response rates were significantly increased. Although the validity coefficients were higher in the experimental conditions, the differences were not statistically significant. However, the validity gains may be practically significant in organizational settings where the goal is to implement the best selection tools available. Given the simplicity of the intervention, I argue that the results are encouraging and more research should be conducted with the goal of refining it further.

Study 2 offered an opportunity to retest Hypotheses 2 and 3 (already tested in Study 1) using a different sample. The results were consistent with Study 1. Across both of the experimental conditions, the reliability of job performance ratings was significantly higher compared to the control condition. Mixed support was found for the differences in the variance of job performance between the control and experimental conditions with only the organizational benefits conditions showing statistically significant higher variance than the control condition.

Study 3 Introduction

Studies 1 and 2 focused on the impact of rater motivation on the validity coefficient and on testing an intervention designed to increase the role rater motivation. Results from these two studies demonstrated that motivated raters generated performance criteria that led to higher validity coefficients. Presumably increases in validity coefficients were due to motivated raters providing more accurate ratings of performance. However, this assumption—increased rater accuracy—was not tested in the previous two studies. Thus, an important issue is whether or not a relationship exists between motivation and accuracy. Rating accuracy is considered one of the most important concerns in performance appraisals (Cronabch, 1955; Murphy & Cleveland, 1995). The purpose of Study 3 is to investigate if rater motivation acts as a moderator in the relationship between objective and subjective measures of job performance. Consistent with the reasoning for the previous hypotheses, motivated raters should be more likely than unmotivated raters to give ratings that accurately reflect the objective performance of their subordinates. As such the following hypothesis is proposed:

Hypothesis 7: Rater motivation will moderate relationships between subjective and objective performance such that the relationship will be stronger when rater motivation is high versus low.

Study 3 Method

Participants and procedure

The participants were 83 managers recruited from the sales division of a Romanian organization. Due to confidentiality concerns raised by the organization, no other sample specific demographic information was available to the author. Data obtained from the organization's HR manager indicates that most of the employees are males (around 70%), around 30 years old and most have a college degree.

Measures

Subjective performance. The managers rated the performance of the salespersons using a five-item scale from Behrman and Perreault (1982). Sample items are "producing a high market share for your company in the territory," "generating a high level of dollar sales," and "exceeding all sales targets and objectives for the territory during the year" on 7-point response format ranging from "1 = *Poor"* to "7 = *Outstanding"*.

Objective performance. Objective performance indicators based on sales were collected from organizational records. Twelve performance levels were created after adjusting for territory potential, workload, company presence in the territory, local economic conditions, and competitors. The algorithm used to derive the performance levels is proprietary and was not disclosed to me by the participating organization.

Rater motivation. Rater motivation was measured with a five-item scale developed by Hedge and Teachout (2000). Example items are: "I care how accurate my

ratings were" and "I made the most accurate ratings I could." Raters responded to these items via a five-point Likert-type response (from "1 = Strongly disagree" to "5 = Strongly agree").

Study 3 Results and Discussion

Means, standard deviations, reliabilities and inter-correlations are presented in Table 7. Objective and subjective performance were positively correlated: $r = .46$, $p < .01$.

Table 7

Descriptive and Correlational Information for Study 3

	M	SD	1	2	3
1. Subjective performance	14.90	6.02	*(.89)*		
2. Objective Performance	5.88	2.44	.46**	*(NA)*	
3. Rater motivation	10.47	5.02	.12	-.01	*(.90)*

** $p < .01$, $N = 83$

Hypothesis 7 proposed that rater motivation will moderate relationships between subjective and objective performance such that the relationship will be stronger when rater motivation is high versus low. This hypothesis was tested using hierarchical moderated multiple regressions (Cohen, Cohen, West, & Aiken, 2003). The interaction term was created by multiplying the predictor variable (subjective performance) and the moderating variable (rater motivation). I then conducted a hierarchical regression

analysis by regressing objective performance on the predictor variable in Step 1, rater motivation in Step 2, and finally the interaction term in Step 3. The interaction term explained a significant amount of variance in objective performance: F (1, 79) = 5.23, p < .05, ΔR^2=0.04. Simple slope tests show that the relationship between subjective performance and objective performance is stronger when rater motivation is high (β = .61, p < .001) compared to when rater motivation is low (β = .25, p > .05). The full results of the regression analyses are presented in Table 8. A plot of the interaction is illustrated in Figure 3.

Table 8

Rater Motivation as Moderator in Study 3

Step	Independent variable	*B*	*SE B*	95% CI	β	ΔR^2
1	Subjective performance	.19	.04	.11, .26	.46**	.21**
2	Rater motivation	-.03	.05	-.13, .06	-.07	.00
3	Subjective performance x Rater motivation	.02	.01	.00, .07	.71**	.04*

Note: N = 83, CI = confidence interval. * p < .05, ** p < .01.

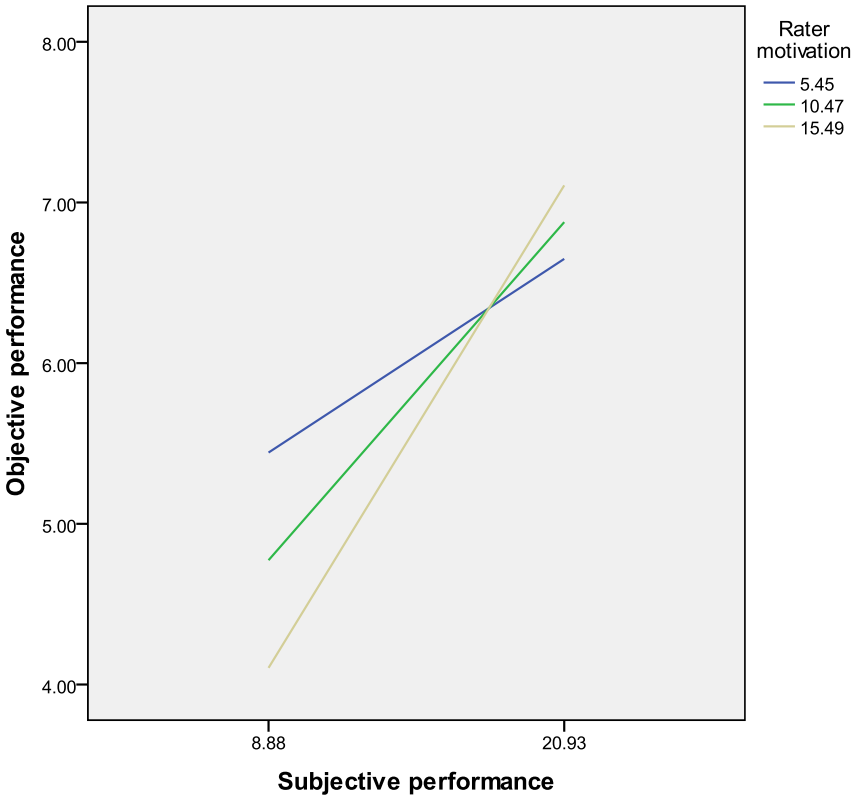

Figure 3: Rater Motivation as Moderator in Study 3

The results of Study 3 suggest that more motivated raters are better at matching objective and subjective performance ratings. There was a strong association between objective and subjective performance for raters high on motivation, while there was no relationship for raters low on motivation. Motivated raters appear to be more accurate; and less likely to fall prey to biases as they seem to be more able to focus on the actual performance of their employees.

General Discussion

The goal of the current paper was to investigate, across three studies and five samples, the role played by rater motivation in personnel selection validation studies. In Study 1, across two samples, I found that rater motivation, measured both directly (self-reported by the raters) and indirectly (the amount of time spent on-line rating), impacted the validity coefficients for two different types of predictors (a cognitive ability test and a job knowledge test). The validity coefficients were significantly higher when rater motivation was high as opposed to low. Another hypothesis fully supported across both of the Study 1 samples concerned the reliability of job performance ratings: motivated raters provided ratings that were more internally consistent than unmotivated raters. Similar results were found when the reliability hypothesis was retested in Study2 using a different sample. High reliability is desirable particularly since the criterion space was narrowly conceptualized and measured in all of the samples used in the current paper (as requested by the participating organizations). No support was found for the hypothesis concerning the variance in ratings; although the raters high on motivation had higher variance in their ratings tan those low on motivation, the differences were not statistically significant. Mixed findings in terms of the variance of the job performance ratings were found when re-testing the hypothesis in Study2 using a different sample. The results of the Study 1 highlight the importance of rater motivation in validation studies and led to Study 2 where I tried to develop, implement and test a short, inexpensive and easy

theoretically-based intervention designed to increase rater motivation, rater response rates, and the validity coefficients. The intervention consisted of manipulation of the instructions such that the benefits of rater participation in validation studies were presented. Two experimental conditions were used: one focusing on the benefits for the individual, the other focusing on the benefits for the organization. Both of the conditions had positive effects in terms of increased response rates and increased rater motivation. Also, the validity coefficients were higher in both of the experimental conditions compared to the control condition, although the differences were not statistically significant. However, I argue that this is a case where small effects are practically significant even though they are not statistically significant. Prentice and Miller (1992) discuss two situations when even small effects can be impressive: when there are minimal manipulations of the independent variable and when the dependent variable is "difficult-to-influence." Both situations seem to apply here: the manipulations were minor and increasing the validity coefficient of the predictors is a difficult task.

As a further testament of the importance of rater motivation, in Study 3 I found that there is a stronger association between objective performance (sales) and subjective ratings for raters high in motivation. Thus, it does appear that raters who are more motivated tend to provide more accurate performance ratings. This finding is important because it dispels any criticism that the observed increases in rating variance in Studies 1 and 2, and by extension the increases in validity coefficients in the two studies, were artifactual in nature. Rather, increasing rater motivation led to an increase in rater accuracy.

The present study makes several contributions to the literature. Previous research on rater motivation has mostly ignored to measure it directly and tended to focus on the differences between ratings made for research purposes and ratings made for administrative purposes. Also, although the importance of validation studies cannot be overstated, very little research has focused on the conditions surrounding validation studies. Similarly, the previous interventions proposed in the literature, such as teams of raters, discussions between raters and rates, have limited use in the case of validation studies where the goal is to obtain fast quality ratings.

Several directions for future research are suggested. First, the role of rater motivation in validation studies can be examined using other predictors, such as personality measures. Second, the intervention suggested here can be compared with other types of interventions (e.g., providing incentives, e.g., Salvemini, Reilly, and Smither, 1993). Third, the role of individual differences in rater motivation should be examined further. In the current study, individualism and collectivism had no impact and the scales used had very low reliabilities (even though I have used them in previous studies and achieved internal consistencies levels in the .70-.80 range). For example, the relationship between rater conscientiousness and rater motivation can be examined. Persons high in conscientiousness are characterized by being responsible, organized, and dependable (e.g., John, Naumann, & Soto, 2008), and thus presumably more likely to be more motivated when they engage in rating their subordinates. Similarly, need for cognition (Cacioppo, Petty, & Kao, 1984) should be examined as a possible antecedent of rater motivation. Persons high in need for cognition are more likely to engage in tasks that are cognitively demanding (i.e., are more likely to use explicit processing). Fourth,

future research should use the recent multidimensional conceptualization of the criterion space (organizational citizenship behaviors and counterproductive work behaviors in addition to task performance, Sackett & Lievens, 2008). Fifth, more research is needed on the role of rater motivation for ratings made for administrative studies. For example, are more motivated raters perceive as more fair? Sixth, the nomological network of rater motivation should be further explored by identifying personal and situational correlates and boundary conditions. Seventh, the simultaneous role played by rater and ratee motivation should be examined. It is possible that further validity gains can obtained by increasing ratee motivation. Similar interventions designed to increase ratee motivation should be tested, as the literature seems to be even more lacking than in the case of rater motivation.

The studies presented in the current paper also have several strengths. First, the data collection was embedded in actual validation studies, thus increasing the external validity of the results. Second, several hypotheses were tested multiple times increasing our confidence in the generalizability of the results to other samples or settings. Third, different conceptualizations of the predictor variables and different ways of measuring rater motivation were used.

The practical implications are straightforward. Rater motivation should be carefully considered when conducting validation studies. Higher validity coefficients are desirable for legal and practical purposes (i.e., establishing cut-off scores). The intervention presented in Study 2 can be used, as one possible way of increasing rater motivation. It's important that the managers and the employees understand the

importance of validation studies. Poor ratings may also be the consequence of larger organizational issues (Harris, Ispas, & Schmidt, 2008).

Several limitations should also be acknowledged. First, all the samples were collected in Romania from Romanian organizations. Although I have no reason to believe that culture played a role in the results obtained, this may limit the generalizability of the findings. Second, although I discussed Harris' (1995) determinants of rater motivation, none of them were measured in the current study. Third, the sample size for Studies 2 and 3 was relatively small. Future research should attempt to replicate the current results using larger, more representative samples. Fourth, rater's previous experience and performance related training should be controlled for in future studies. It is possible that more experienced, more trained raters will be more motivated; however I was not able to test this hypothesis in the current study. Fifth, the use of the sales index as a measure of objective job performance can be criticized as objective measures are prone to their own biases and error.

In conclusion, the current series of studies show that rater motivation plays an important role in validation studies and should be included in future theoretical and practical models of performance ratings.

References

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695–716.

Austin, J. T, & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77,* 836-874.

Bardos, A. (2003). The General Ability Measure for Adults. In R. S. MacCallum (Ed.), *Handbook of nonverbal assessment* (pp. 163-174). New York, NY US: Kluwer Academic / Plenum Publishers.

Barrett, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology, 66,* 1–6.

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 of the new millennium: What do we know and where do we go next? *International Journal of Selection & Assessment, 9,* 9-30.

Behrman, D., & Perreault, W. D., Jr. (1982). Measuring the performance of industrial salespersons. *Journal of Business Research, 10,* 355–370.

Borman, W.C. (1975). Effects of instruction to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology, 60,* 556-560.

Borman, W.C. (1977). Consistency of rating accuracy and rating error in the judgment of human performance. *Organizational Behavior and Human Performance, 20,* 238-252.

Brannick, M. T., Levine, E. L., & Morgeson, F. P. (2007). *Job analysis: Methods, research, and applications for Human Resource Management* (2nd Ed.). Thousand Oaks, CA: Sage Publications.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306-307.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, B. W. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in the need for cognition. *Psychological Bulletin, 119,* 197-253.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74,* 130-135.

Cohen, J., Cohen, P., Aiken, L. S., & West, S. G. (2003). *Applied multiple regression – correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Crawford, N., Festa, T., Sutton, L., & Bardos, A. N. (1999). *The predictive validity of the General Ability Measure for Adults*. Paper presented at the annual meeting of the National Association of Neuropsychology, San Antonio, Texas.

Cronbach, L. (1955). Processes affecting scores on understanding of others and assumed similarity. *Psychological Bulletin, 52,* 177-193.

DeCotiis, T. & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. *Academy of Management Review, 3,* 635-646.

Educational Testing Service. (1973). *Administrator's guide for the cooperative school and college ability tests (SCAT)*. Princeton, NJ: Author.

Evans, J. St. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128,* 978-996.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An *EPM* guidelines editorial. *Educational and Psychological Measurement, 6,* 517-531.

Fiske, S., & Taylor, S. (1984). *Social cognition*. Reading, MA: Addison-Wesley.

Gatewood, R.D., & Feild, H.S. (2005). *Human resource selection* (7th ed.). FortWorth, TX: Harcourt Brace.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-salience theory of survey participation: Description and illustration. *Public Opinion Quarterly*, 64, 299-308.

Hakstian, A. R.,& Whalen, T. E. (1976).A k-sample significance test for independent alpha coefficients. *Psychometrika, 41*, 219-231.

Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management, 20,* 737-756.

Harris, M.M., Ispas, D., & Schmidt, G. F. (2008). Inaccurate performance ratings are a reflection of larger organizational issues. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 190-193.

Harris, M. M., Smith, D. E., & Champagne, D. (1995). A field study of performance appraisal purpose: Research- versus administrative-based ratings. *Personnel Psychology, 48,* 151-160.

Hedge, J. W., & Teachout, M. S. (2000). Exploring the Concept of Acceptability as a

Criterion for Evaluating Performance Measures. *Group and Organization Management,*

25, 22-55.

Ilgen, D. R., & Knowlton, W. A., Jr. (1980). Performance attributional effects on feedback from superiors. *Organizational Behavior and Human Performance, 25*, 441–456

Iliescu, D., & Livinţi, R. (2008). *Romanian Manual for GAMA – General Ability Measure for Adults.* Cluj-Napoca, Romania: Odiseea.

Ispas, D., Iliescu, D., Ilie, A., & Johnson, R.E. (2010). Examining the criterion related validity of the General Mental Ability Measure for Adults: A two sample investigation. *International Journal of Selection and Assessment, 18*, 224-227.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty*: *Heuristics and biases*. Cambridge, MA: Cambridge University Press.

Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman Brief Intelligence Test.* Circle Pines, MN: American Guidance Service.

Jackson, D. N. (2003). *Manual of the Multidimensional Aptitude Battery – II.* London: Sigma Assessment Systems.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research,* (pp. 114-158). New York, NY: Guilford Press.

Johnson, R. E., Rodopman, O. B., Akirmak, U., Gray, A., & Ispas, D. (2010). *Applying dual processing theory to organizational behavior: Organizational justice as an example*.

Jawahar, I.M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology*, *50,* 905-925.

Judge, T. A., & Erez, A. (2007). Interaction and intersection: The constellation of emotional stability and extraversion in predicting performance. *Personnel Psychology*, *60*, 571-594.

Kanfer, R. (1990). Motivation theory and Industrial/Organizational psychology. In M. D. Dunnette and L. Hough (Eds.), *Handbook of industrial and organizational psychology. Volume 1. Theory in industrial and organizational psychology* (pp.75-170). Palo Alto, CA: Consulting Psychologists Press.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72-107.

Levy, P. E., & Willliams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management, 30,* 881–905.

Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, *58,* 259–289.

Longnecker, C.O., Gioia, D.A., & Simes, H.P. (1987). Behind the Mask: The Politics of Employee Appraisal. *Academy of Management Executive, 1,* 183-193.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

Maier, M. H. (1988). On the need for quality control in validation research. *Personnel Psychology*, *41,* 497−502.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb,W. L., III (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60,* 63−91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology, 86,* 730−740.

Mero, N. P. & Motowidlo, S. J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology, 80,* 517-524.

Muchinsky, P. M. (2004). When the psychometrics of test development meets organizational realities: A conceptual framework for organizational change, examples and recommendations. *Personnel Psychology, 57,* 175–209.

Murphy, K. R., & Cleveland, J. N. (1991). *Performance appraisal: An organizational perspective.* Needham Heights, MA: Ally & Bacon

Murphy, K. R., & Cleveland, J. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: Sage.

Naglieri, J. A., & Bardos, A. N. (1997). *General Ability Measure for Adults*. Minneapolis, MN: Pearson Assessments.

Napier, N. K., & Latham, G. P. 1986. Outcome expectancies of people who conduct performance appraisals. *Personnel Psychology, 39,* 827-837.

Oppenheimer, D., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867-872.

Petty, R.E., & Cacioppo, J.T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology, 37*, 1915-126.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123-205.

Petty, R. E., Wegener, D. T., & Fabrigar, L. R. (1997). Attitudes and attitude change. *Annual Review of Psychology, 48,* 609–647.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*, 160-164.

Rand, T. M., & Wexley, K. N. (1975). Demonstration of the effect "similar to me" in simulated employment interviews. *Psychological Reports, 36*, 535–544.

Roch, S. G. (2007). Why convene rater teams: Investigation of the benefits of anticipated discussion, consensus, and rater motivation. *Organizational Behavior and Human Decision Processes, 104*, 14-29.

Russell, C.J., Settoon, R.P., McGrath, R., Blanton, A.E., Kidwell, R.E., Lohrke, F.T., Scifries, E.L., & Danforth, G.W. (1994). Investigator characteristics as moderators of selection research: A meta-analysis. *Journal of Applied Psychology, 79*, 163-170.

Sackett, P.R., & Lievens, F. (2008). Personnel selection. *Annual Review of Psychology, 9*, 419-45.

Salvemini, N. J., Reilly, R. R., & Smither, J. W. (1993). The influence of rater motivation on assimilation effects and accuracy in performance ratings. *Organizational Behavior and Human Decision Process, 55*, 41–60.

Satpute, A. B., & Lieberman, M. D. (2006). Integrating automatic and controlled processes into neurocognitive models of social cognition. *Brain Research, 1079,* 86–97.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407–422.

Shipley, W. C. (1991). *Shipley Institute of Living Scale*. Los Angeles: Western Psychological Services.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning.* Hillsdale, NJ: Erlbaum.

Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin.* Chicago: The University of Chicago Press.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8,* 220-247.

Triandis, H.C., & Gelfand, M.J. (1998). Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology, 74,* 118-128.

Uniform guidelines on employee selection procedures. (1978). 43 Fed. Reg. 38, 290 – 38,315.

Wexley, K.N., & Klimoski, R. (1984). Performance appraisal: An update. In K.M. Rowland & G.D. Ferris (Eds.), *Research in personnel and human resources management*. Greenwich, CT: JAI Press.

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society (Series B)*, *21*, 396–399.

Wonderlic Inc. (2002). *Wonderlic Personnel Test and Scholastic Level Exam user's manual*. Libertyville, IL: Wonderlic, Inc.

About the Author

Dan Ispas received a B. A. in Journalism and English and a post-graduate academic degree in social communication and public relations from his native country, Romania. He received his M.A. in Industrial and Organizational (I/O) Psychology with a minor in Statistical Methods from the University of South Florida (USF) in 2008. During his doctoral studies at USF, he was lead instructor for the following courses: *Tests and Measures, Psychological Statistics,* and *Research Methods in Psychology*. In 2008 he completed an I/O psychology internship at Procter & Gamble. His research was published in *Human Resource Management Review, Industrial and Organizational Psychology: Perspectives on Science and Practice, International Journal of Selection and Assessment, Pastoral Psychology* and *Psihologia Resurselor Umane*. Between 2007 and 2010, he presented 20 papers and chaired 2 symposia at the SIOP and Academy of Management Annual Conferences. At the 2010 SIOP Annual Conference he was awarded, as principal investigator, a SIOP Small Grant and a paper based on his master's thesis was a finalist for the John C. Flanagan Award and recognized as a Top Poster.