When Rating Format Induces Different Rating Processes: The Effects of Descriptive and Evaluative Rating Modes on Discriminability and Accuracy

Laurent Cambon, Dirk D Steiner

▶ **To cite this version:**

Laurent Cambon, Dirk D Steiner. When Rating Format Induces Different Rating Processes: The Effects of Descriptive and Evaluative Rating Modes on Discriminability and Accuracy. Journal of Business and Psychology, 2015, 30 (4), pp.795 - 812. 10.1007/s10869-014-9389-y . hal-01881712

HAL Id: hal-01881712
https://hal.univ-cotedazur.fr/hal-01881712

Submitted on 26 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Running head: MODE OF RATING, DISCRIMINABILITY, AND ACCURACY

When Rating Format Induces Different Rating Processes: The Effects of Descriptive and Evaluative Rating Modes on Discriminability and Accuracy

Laurent Cambon and Dirk D. Steiner

Laboratoire d'Anthropologie et de Psychologie Cognitive et Sociale, Université de Nice-Sophia Antipolis, France

Please send correspondence to:

Laurent Cambon
Laboratoire d'Anthropologie et de Psychologie Cognitive et Sociale
Université de Nice-Sophia Antipolis
24 avenue des Diables Bleus
06357 Nice cedex 4 France
E-mail: cambon@unice.fr
Phone : ++33 489 881 461

Abstract

Purpose -- We examined how different kinds of rating formats, and their interaction with purposes of rating (administrative vs. developmental), induced different performance rating processes and their consequences for rating accuracy.

Design/methodology/approach -- In two experiments, participants rated seven targets presented via videotapes using modes of rating giving access to: a) descriptive knowledge (rating scales were a target's observable behaviors: Descriptive Behavior--DB), b) evaluative knowledge (rating scales were others' behaviors that the target tended to afford: Evaluative Behavior--EB), or c) a mix of the two knowledge types (rating scales were traits). Indexes of discriminability (within- and between-ratee discriminability) and of accuracy (differential elevation and differential accuracy) were collected.

Findings – The results showed that EB rating scales led to higher between-ratee discriminability and differential elevation than other modes of rating, whereas DB rating scales led to higher within-ratee discriminability than the other modes.

Implications – Our results indicate that EB rating scales are more suited to comparing different ratees (e.g., an administrative purpose for rating), whereas DB scales are more suited to identifying strengths and weaknesses of a particular ratee (e.g., a developmental purpose).

Originality/value – Our experiments are the first to apply dual-knowledge (descriptive vs. evaluative) theory to a performance appraisal context and to examine rating purpose in interaction with these two forms of person knowledge. The results, consistent with theoretical predictions, indicate that using rating scales with different types of content as a function of the rating purpose will produce more appropriate performance ratings.

Key words: evaluative knowledge, descriptive knowledge, performance appraisal, discriminability, rating format, accuracy, purpose of evaluation.

When Rating Format Induces Different Rating Processes: The Effects of Descriptive and Evaluative Rating Modes on Discriminability and Accuracy

Various professional activities are based on evaluative judgments: personality assessment, employee job performance ratings, appraisal interviews and feedback, employment interviews, and 360° assessments are some examples of these judgment contexts. For objectification and communication reasons, these judgments are often reported on rating forms. In organizational psychology, numerous studies have been conducted on the impact rating format, or mode of rating, has on evaluative judgments, and most notably in the context of performance appraisal. However, after reviewing a large body of research on a variety of formats, Landy and Farr (1980) concluded that there was no unequivocal evidence favoring one rating format over others. They therefore called for a moratorium on rating format research and for a shift in focus of performance appraisal research toward the social-cognitive processes involved in rating. The moratorium on rating format research has largely been respected over the past 30 years (Borman, Bryant, & Dorio, 2010; however, see Goffin & Olson, 2011 and Tziner & Kopelman, 2002 for continuing lines of research on rating formats), and many advances have been made on the underlying processes of performance appraisal (Murphy & Cleveland, 1991) as well as on the effective use of appraisal information in developing employees, among other topics.

Despite the many advances accumulated through this vast body of research, examining rating formats may still have the potential of making contributions to the evaluative judgment literature, and especially to the performance appraisal field, particularly if studied in light of some other relevant advances in the literature, notably regarding the social-cognitive processes. Thus, in this paper, we review research in social cognition that shows how descriptive versus evaluative knowledge may be more compatible with certain types of

behaviorally based rating scales. We also examine the interaction of these two types of knowledge with two different categories of traits useful in predicting performance, namely agency and communion. Further, we discuss how these types of knowledge relate to the different forms of accuracy described by Cronbach (1955). Then, we explore how these knowledge types relate to the administrative versus developmental purposes of performance rating. Finally, we present results from two experiments showing the advantages of using evaluatively-based rating scales for administrative purposes and descriptively-based scales for developmental purposes. Although our contribution is contextualized in the performance appraisal process, its conclusions seem relevant for the more general problem of evaluating others in terms of performance or personality (for instance, in various applications of individual assessment or performance improvement, through appraisal interviews, selection interviews, or personality assessment).

Rating Formats and Rating Processes

Despite the relative abandon of rating format research in favor of emphasis on the social-cognitive processes involved in rating that followed Landy and Farr's (1980) landmark publication, it has been suggested repeatedly that rating scale formats can influence the processes, beginning with attention to and observation of behavior (McDonald, 1991; Murphy & Constans, 1987; Murphy & Pardaffy, 1989; Piotrowski, Barnes-Farrell, & Esrig, 1989), continuing with storage of performance information in memory (DeNisi, Robbins, & Summers, 1997), and ending with recall and evaluation of performance information (Fay & Latham, 1982; Heneman, 1988; Murphy & Constans, 1987; Murphy, Martin, & Garcia, 1982; Murphy & Pardaffy, 1989; Piotrowski et al., 1989; Pulakos, 1984). Thus, the possibility of format-related differences in ratings may be worthy of reconsideration if we examine the relation between rating formats and the social-psychological processes they induce. More precisely, based on a model of the functions of psychological traits, the theory of dual-

knowledge (Beauvois & Dubois, 2000), we will argue that different rating formats should serve different functions in performance appraisal and more generally in psychological assessment. Indeed, psychological traits often constitute the basis for items included in rating scales, even if behaviorally-based rating approaches have been preferred for a number of years (e.g., Landy & Farr, 1980; 1983; Latham & Wexley, 1977; Murphy & Cleveland, 1991; Smith & Kendall, 1963; Woehr & Roch, 2012). In many types of evaluation (i.e., graphic rating scales), traits are used as scale anchors; in more specific appraisal methods (i.e., based on critical incidents of behavior such as in Behaviorally Anchored Rating Scales, Smith & Kendall, 1963; or Behavioral Observation Scales, Latham & Wexley, 1977), traits could be conceived as the psychological constructs the behaviors operationalize, as we will show more clearly in the following paragraphs. Thus, studying the psychological function of traits could give insight into various psychological assessment processes.

Descriptive and Evaluative Functions of Psychological Traits

Traditionally, in personality theories (Goldberg, 1993; John, 1990) and in the social cognition literature (Bassili, 1989; Srull & Wyer, 1979), personality traits were considered to be descriptive, that is, lexical units summarizing behavioral consistencies that people observed during their interactions with others. Thus, describing someone as "honest," means that one has observed the person turn in a wallet found in the street. Opposed to this descriptive conception of traits is their evaluative function which has its roots in the Gibsonian theory of perception (e.g., McArthur & Baron, 1983; Zebrowitz & Collins, 1997). Gibsonian theory suggests that the environment is perceived in terms of its functional properties rather than its descriptive qualities. These functional properties involve affordances, that is, opportunities to act on objects. Applied to person perception, this means that traits are conceived of as affordances, that is, as a set of properties related to an object's

utility for a judge in a given situation. For example, a person's agreeability refers directly to that person's utility for me at the time when I interact with him/her.

The descriptive and evaluative conceptions of traits have been reconciled by Beauvois and Dubois (2000) in their dual-knowledge theory (descriptive and evaluative knowledge) which postulates that a trait is both a descriptive tool and a generalized affordance. Indeed, Beauvois and Dubois (2000) have shown that a trait (for example "honest") is connected both to a descriptive statement ("someone who always admits his own mistakes") and to an affordance ("someone people go to for an objective opinion ") at the semantic level. The first type of linkage is referred to as a descriptive behavior (DB), that is, a characteristic of the target. The second type of linkage is considered an evaluative behavior (EB) because it is a behavior directed toward the target based on an evaluation of what one can do with the target, and it says nothing about the person's descriptive characteristics. More particularly, Beauvois and Dubois (2000) showed that EBs are more accessible when one is dealing with the most evaluative traits (i.e., honest), whereas DBs are more accessible when one is dealing with the least evaluative traits (i.e., nervous). Moreover, Dubois and Tarquinio (1998) showed that professional raters recalled and used EBs more than DBs when evaluating a ratee.

Mignon and Mollaret (2002) went a step further. Using a zero acquaintance paradigm, they asked participants to watch and judge targets each presented by an 8 second silent video exposing a sequence of behaviors. The judgments were reported on: traits, DBs or EBs strongly associated with the traits. They reasoned and showed that descriptive knowledge should lead to differentiating between several characteristics of a target (increasing within-ratee discriminability), whereas evaluative knowledge should lead to differentiating between targets, which one is the best and which is the worst (increasing between-ratee discriminability). Thus, using DBs as rating scales should lead participants to more within-

ratee discriminability than with either trait or EB rating scales whereas using EB should lead them to more between-ratee discriminability than with either traits or DB.

Although obtained in the context of every day social interactions and not in a performance appraisal situation, these results could have interesting implications for performance appraisal practices and other assessments in organizations because within-ratee and between-ratee discriminability correspond to two fundamental purposes of evaluation, namely the developmental and the administrative ones[1] (Cleveland, Murphy, & Williams, 1989; DeNisi & Peters, 1996; Murphy & Cleveland, 1991). Administrative purposes include ratings for merit increase, hiring and promotion decisions, disciplinary action, and retention decisions--all actions which focus on distinguishing between ratees. The developmental purpose concentrates on feedback to ratees to assist in their development and thus focuses on distinguishing different characteristics within each ratee. Therefore, the results of Mignon and Mollaret (2002) suggest that a rating format using EBs as rating scales should improve appraisals requiring comparing ratees, as with administrative purposes, whereas a rating format using DBs should improve appraisals in which a person's relative strengths and weaknesses must be identified, as with the developmental purposes of performance appraisal.

Although DB and EB have never been used in the performance appraisal literature, one could find partial support for our suggestions in the literature opposing specific to global rating formats. Indeed, DB, by its focus on target actions, is more akin to specific rating formats such as Behaviorally Anchored Rating Scales (BARS; Smith & Kendall, 1963) or Behavioral Observation Scales (BOS; Latham & Wexley, 1977). On the contrary, by focusing on abstract expectations toward a target, EB is reminiscent of global rating scales such as graphic rating scales. Although, no clear advantage in favor of one format (global vs. specific) over the other has emerged in the literature (Fay & Latham, 1982; Heneman, 1988), the specific formats are generally considered superior to other formats in fostering developmental

purposes because they define job performance in relatively objective terms which could lead to setting more specific goals (Tziner & Kopelman, 2002). On the other hand, global formats are theorized as more appropriate for administrative purposes because they are composed of more evaluative and broader constructs which are pertinent for ranking ratees or predicting future performance (Bernardin & Beatty, 1984). However, before considering the usefulness of the results of Mignon and Mollaret (2002) for performance appraisal and other evaluative practices, it is necessary to test the discriminability and accuracy of these two rating modes (EBs and DBs) in a context which is more closely related to that of a professional appraisal. Thus, the main objective of the studies presented here is to test these two rating modes and Mignon and Mollaret's hypotheses using a video observation paradigm, that is, a paradigm presenting long sequences of behaviors. Such a paradigm allows for simulating processes much more in line with those implied in real performance appraisal situations. Although our main interest is with EBs and DBs, we included traits in our hypothesis as a reference point from which we can compare the effect of the two types of rating formats and because traits constitute the base on which the EBs and DBs are constructed. Thus, the first hypothesis is:

> *Hypothesis 1: A rating mode based on DBs should produce more within-ratee discriminability than a mode based on either EBs or traits; conversely, a rating mode based on EBs should produce more between-ratee discriminability than a rating mode based on either DBs or traits.*

Dimensions of Performance

Mignon and Mollaret (2002) examined the interaction between rating mode and judgments of communion and agency because these dimensions were based directly on interpersonal relations and as such were relevant for the study of affordances. The communion dimension refers to interpersonal behaviors related to socialization and friendship, whereas the agency dimension refers to interpersonal behaviors related to power

and personal growth. While these two dimensions are rarely used in Work and Organizational psychology, they could be related to the well-known Big Five dimensions. For example, Communion can be associated to the dimensions of Agreeableness, Neuroticism, and the sociability facet of Extraversion, whereas Agency matches with the content of Conscientiousness, Intellect, and the ambition facet of Extraversion (De Young, Quilty, & Peterson, 2007). Based on past results (Funder & Dobroth, 1987; Paulhus & Reynolds, 1995), Mignon and Mollaret (2002) showed that between-ratee discriminability was much more accentuated on the agency dimension than on communion. Although they did not propose an explanation of this result, we suggest that agency, more than communion, deals with characteristics that are useful for attributing social and professional value to people (Cambon, Djouary, & Beauvois, 2006; Wojciszke, Abele, & Baryla, 2009). Thus, this dimension is probably more relevant for ranking people (i.e., between-ratee discriminability) in a performance appraisal situation. On the basis of these past results, we hypothesized that:

Hypothesis 2: Ratings made on the agency dimension should produce more between-ratee discriminability than ratings made on the communion dimension.

Rating Accuracy

Our focus on between- and within-discriminability is determined by our interest in the processes at the core of appraisal practices. We thus have hypothesized that different modes of rating (EB vs. DB) should lead raters to focus their attention on different kinds of information. Of course, the activation of a process has nothing to do with its efficiency: whether more or less within- or between-discriminability occurs does not address the accuracy of the process. Thus, the third objective of this paper is to determine which type of rating mode, if any, leads to greater accuracy. We propose that the different modes produce different forms of accuracy. More specifically, the judgmental accuracy of performance ratings using EB versus DB versus Trait modes of formulating rating scales will be assessed

with the four distinct forms of accuracy developed by Cronbach (1955): elevation (EL), the accuracy of ratings, averaging across all ratees and performance dimensions; differential elevation (DEL), accuracy in discriminating among ratees, averaging across performance dimensions; stereotype accuracy (SA), accuracy in discriminating among performance dimensions, averaging across ratees; and differential accuracy (DA), accuracy in discriminating among ratees, within each performance dimension.

We make predictions regarding the specific form of accuracy that will be enhanced by each rating mode by comparing the logic of the different components of accuracy with that of the within- and between-ratee discriminability indexes. We conclude that there seems to be no link between the two discriminability indexes and EL and SA (there is no between-ratee or within-ratee discriminability implied in the calculation of EL and SA). However, between-ratee discriminability and DEL are close measures as they both examine how raters discriminate among ratees. Further, DA includes a component of between-ratee discriminability (discrimination among ratees) as well as a component assessing within-ratee discriminability (discrimination among ratees within each performance dimension) in its calculation. Thus, because EBs increase between-ratee discriminability and DBs within-ratee discriminability, one could hypothesize that the EB rating mode should produce better DEL scores than the DB and Trait modes, whereas the DB and EB modes should produce better DA scores than traits[2]

As no clear predictions could be made for the other two components of accuracy, EL and SA, we will not examine them further in this research. We therefore formulated hypotheses only for DEL and DA:

Hypothesis 3: EBs should give rise to more DEL than DBs and Traits; conversely, DBs and EBs should give rise to more DA than Traits.

Given that (a) between-ratee discriminability and DEL are close measures, based on related algorithms, and (b) ratings made on the agency dimension should produce more between-ratee discriminability than ratings made on the communion dimension (Hypothesis 2), by deduction, it is also possible to predict that:

Hypothesis 4: Agency ratings should give rise to more DEL than communion ratings.

These hypotheses will be tested in two experiments. The first one will examine these four hypotheses, then a follow-up study (Experiment 2), will also test complementary hypotheses in line with the administrative or developmental purposes of performance appraisal that will be introduced later. We tested these hypotheses in the situation of an appraisal interview. An appraisal interview generally involves an interaction between a ratee and a rater during which performance information is reviewed. The rater may present evaluation information and ask the ratee questions about performance during the review period. The appraisal interview situation is relevant for the study of performance appraisal processes, but it also can shed light on other evaluative practices such as employment interviews and personality assessment. In order to make these generalizations possible, we structured the interaction between the rater and the ratee so as to minimize the interventions of the rater and to maximize those of the ratee. So, from the point of view of participants, this situation was very similar to that of observing real performance except that performance itself was not shown but declared by ratees.

Experiment 1

Method

Development of Rating Scales

A series of preliminary studies was conducted to select the traits, EBs, and DBs used in developing the rating scales for the two experiments (for details of this procedure, see Mignon & Mollaret, 2002). The aim was to select EBs and DBs that were equally

representative of the traits. Thus, based on our preliminary studies, we selected three traits for each of the four categories formed by crossing the two dimensions of evaluation, communion and agency, with the negative and positive valence of the dimensions (positive communion, negative communion, positive agency, and negative agency). The traits were selected because they were: (a) widely used in the appraisal practices of managers, and (b) easy to express in the scripts for the different targets (see videotapes section). Then, one EB and one DB were selected to illustrate each of these 12 traits. These behaviors[3] were selected during two different pretests. The first asked participants (N = 100) to (a) choose the one behavior among five EBs or five DBs that was most representative of the trait, and (b) rate the prototypicality of each EB or DB chosen. We retained the most frequently chosen and prototypical behaviors. To verify further that DBs and EBs were trait specific, we conducted a second pretest asking undergraduate students (N = 80) to perform an association task between the behaviors (DBs or EBs) and the traits. The links made by these participants were compared with the results of the first pretest. An association was considered to be "correct" when a participant made the expected choice. Correct associations were obtained at a high level of frequency (for DBs, 82% correct associations on average; for EBs, 78% correct on average). Combined, these series of tests show that the EBs and DBs of this material were similarly trait-specific. Examples of the traits and their corresponding EBs and DBs are presented in Table 1.

To construct the three rating scales for the appraisal of managers in our experiments, we used the twelve traits, EBs, or DBs as items representing the four dimensions. Ratings used a Likert-type format ranging from *absolutely did not characterize the manager* (1) to *absolutely characterized the manager* (7). Items that depicted ineffective performance were rescaled so that high values indicated high performance.

Development of Videotapes

We developed scripts for seven videotaped appraisal interviews of managers. Each videotape illustrated one manager's interview; during the interview, all managers responded to the same questions about their work. The questions (e.g.: how can you explain the results you obtained this year?) and the responses were scripted in order to provide information on the performance level of the manager on the twelve traits of the rating scales. The responses illustrated different performance levels: (a) among the seven managers, and (b) for each manager, among the different rating scales. Seven drama students from a nearby university were each filmed in close-up playing the written role of a manager. The videotapes were from 5 to 7 minutes long (Murphy, Garcia, Kerkar, Martin, & Balzer, 1982). Two pilot studies confirmed the successful manipulation of the performance levels within and between managers[4].

Participants and Procedure

Sixty third-year management students at a French university who were familiar with the appraisal interview participated in the experiment in small groups (two to four per group) on a voluntary basis. No incentives were given for participation. First, we presented and explained the use of the rating format. Then, participants viewed the seven videotaped appraisal interviews of supposed managers in a toy factory, rating each one immediately after viewing the tape using the evaluation forms provided at the beginning of the experiment. Four different randomly determined presentation orders of the seven videotapes were used to control for possible order effects. Thus, the participants were randomly assigned to one of twelve different experimental groups obtained by crossing three modes of rating (Trait scales, EB scales, DB scales) with four videotape presentation orders.

Dependent Variables

Discriminability measures. Using the ratings obtained with each rating mode (traits vs. EBs vs. DBs), we computed two discriminability measures, within- and between-ratee.

Within-ratee discriminability measured variance across the rating dimensions for each ratee. We calculated four scores of within-ratee discriminability, one each for the positive and negative poles of the agency and communion dimensions. Within-ratee discriminability was operationalized (see DeNisi & Peters, 1996) as the average within-ratee standard deviation across the three rating items composing each dimension and standardized within each dimension (consistent with the suggestion of Pulakos, Schmitt, & Ostroff, 1986). Between-ratee discriminability (DeNisi & Peters, 1996) measured the extent to which raters discriminated across ratees. It was operationalized as the average standard deviation of the three rating items across ratees computed for each of the four dimensions.

Accuracy measures. Generally, judgmental accuracy refers to how congruent ratings of performance are with "true" ratings of performance (Murphy, Kellam, Balzer, & Armstrong, 1984). Two of Cronbach's (1955) four components of accuracy were used as the dependent variables representing rating accuracy: (1) Differential elevation (DEL), (2) Differential accuracy (DA).

True scores. We derived the true score measures of performance for the three rating modes by conducting a study of expert raters. Twenty-one managers whose work included conducting performance appraisals, and who had training in social and organizational psychology, served as expert raters. They were thoroughly briefed on the nature and contents of the tapes and were given multiple opportunities to view the tapes and the script outlines prior to rating them (see Borman, 1977). Each expert rated all seven tapes using one mode of rating scale (traits *vs.* EBs *vs.* DBs). Thus, three groups of seven experts were used. The experts' mean rating for each item served as the estimate of the true score of performance for both experiments (the present one and Experiment 2). We computed the intraclass single expert indexes (ICC(C,1)) for each performance dimension and rating method (McGraw & Wong, 1996). The same intraclass indexes were computed for the student raters. They are

presented in Table 2 for Experiment 1 and in Table 3 for Experiment 2. The expert reliability estimates were higher than the student ones in both experiments, thus attesting that expert raters were truly providing expert ratings.

Results

Preliminary Analyses

We first examined whether the targets were rated similarly irrespective of the rating format and the dimension. Thus, we computed a series of correlations using targets as the unit of analysis. As can be seen (Table 4), the correlations among the ratings of the various dimensions using the different rating formats were high, confirming that each target received consistent ratings across dimensions and rating formats. But it should also be noted that the correlations between agency and communion are rather high and suggest a potential confound between these two constructs. Although problematic, it should be stressed that such correlations between agency and communion are frequent in the literature. When examining group and cultural stereotypes, they are often negatively related, but in trait and person judgment, they are often positively related (see Judd, James-Hawkins, Yzerbyt, & Kashima, 2005). In this latter case, correlations of the magnitude of .60 and more are often observed (Abele, Uchronski, Suitner, & Wojciszke, 2008; Montoya & Horton, 2004; Singh et al., 2009).

Discriminability Measures

We submitted the within-ratee and the between-ratee discriminability measures to the same 3 (Mode of Rating: *Traits vs. EBs vs. DBs*) X 2 (Dimension of scales: *communion vs. agency*) X 2 (Valence of scales: *positive vs. negative*)[5] analysis of variance (ANOVA) with the first variable as a between-participants factor and the other two as within-participants factors (see Table 5 for the ANOVA summary table).

Within-ratee discriminability. The only significant effect was the main effect of Mode of Rating. A series of planned comparisons showed, as hypothesized (H1), that DBs significantly produced the greatest within-ratee discriminability (M = 0.73) compared to either EBs (M = 0.48), F (1, 57) = 61.04, $p < .001$, η^2 = 0.52, or Traits (M = 0.58), F (1, 57) = 21.96, $p < .001$, η^2 = 0.30. The difference between Traits and EBs was also significant, F (1, 57) = 9.77, $p < .005$, η^2 = 0.18.

Between-ratee discriminability. As predicted (H1), the main effect of Mode of rating was obtained and showed that EBs produced the greatest between-ratee discriminability (M = 2.23) compared to either DBs (M = 1.97), F (1, 57) = 7.24, $p < .01$, η^2 = 0.16, or Traits (M = 1.85), F (1, 57) = 17.57, $p < .001$, η^2 = 0.28.

In accordance with Hypothesis 2, the main effect of Dimension of scales was significant and revealed that evaluations on agency (M = 2.06) gave rise to more between-ratee discriminability than evaluations on communion (M = 1.97). Unexpectedly, the analysis revealed a main effect of Valence of scales and the interaction implying all variables. Follow-up analyses showed that the difference between agency and communal dimensions was only obtained for the EB (for positive as for negative scales) and the DB modes of rating, but in this latter case, for positive scales only (Table 6).

Between-ratee discriminability has both a true score (real differences between the ratees) and an error component that reflects the idiosyncratic perspectives of raters. Thus, it is possible that the differences observed here in between-ratee discriminability among rating formats reflect greater error between-ratee variance in EB ratings, rather than greater true variance in those ratings. To evaluate this possibility, we used the Variance Components procedure (with the ANOVA method)[6] to fit a model with ratees and raters as two random effects for each rating. We repeated this procedure for each performance dimension X valence X rating format.[7] If, as we hypothesized, it is the true score that is driving higher between-

ratee discriminability for the EB rating format, then ratee variance should be greater than both a) rater or error variance, and b) DB or traits rating format. This was clearly the case. The ratee variance components were higher than the error and rater variance components, for positive communion (.63), negative communion (.75), positive agency (.76), and negative agency (.71). Moreover, these percentages of variance explained by ratees were higher than those obtained for either the DB (.25, .23, .34, .35) or the trait (.31, .29, .48, .45) rating format.

Accuracy Measures

The analyses on the two accuracy measures used the same ANOVA design (Table 5).

Differential Elevation. The ANOVA with DEL as the dependent variable revealed the predicted main effect of Mode of rating showing that EBs (M = .14) gave rise to better accuracy scores than traits (M = .41), F (1, 57) = 109.98, $p < .001$, η^2 = 0.67, and DBs (M = .39), F (1, 57) = 98.33, $p < .001$, η^2 = 0.49. A main effect of Dimension of rating was also significant revealing that the agency dimension (M = .26) gave more accurate scores than communion (M = .37). The interaction between Mode of rating and Dimension of rating and the interaction implying all three variables did not add crucial information (Table 7): for the EB mode of rating, the agency dimension gave rise to significantly better accuracy scores (LSD test) than the communion one on positive scales, this difference was not significant on negative scales.

Differential accuracy. The analysis on DA revealed the predicted main effect of Mode of rating showing that EBs (M = .09) gave rise to better accuracy scores than traits (M = .16), F (1, 57) = 88.53, $p < .001$, η^2 = 0.61, and DBs (M = .12), F (2, 57) = 11.12, $p < .005$, η^2 = 0.16. The difference between the DB and Trait modes was also significant, F (1, 57) = 36.89, $p < .001$, η^2 = 0.39. A Mode of rating X Dimension of rating interaction showed that the difference between the EB and DB modes of rating was only significant (LSD test) for the

communal dimension (M_{EB} = 0.091, M_{DB} = 0.124) but not for the agency dimension (M_{EB} = 0.092, M_{DB} = 0.112).

Discussion

The results of this study reproduced the main results obtained by Mignon and Mollaret (2002) in a performance appraisal context: EB ratings were more efficient in discriminating between targets, whereas DB ratings were more efficient for discriminating dimensions within a target. It should be noted that we have also reproduced Mignon and Mollaret's results concerning the main effect of the dimensions of evaluation: Evaluations on agency produced more between-ratee discriminability than evaluations on communion. This result is noteworthy considering the fact that the strong correlations obtained between these two dimensions could have obscured the effect of agency on between-ratee discriminability. However, we unexpectedly obtained a main effect of the valence of the scales on this discriminability index. Specifically, negative scales gave rise to more between-ratee discriminability than positive ones. At this point, we have no explanation for this effect which is in need of replication. We will examine this again in Experiment 2.

The results related to accuracy showed two patterns. First, the agency dimension produced better accuracy scores than communion but only for the differential elevation component. This effect confirms Hypothesis 4. The most surprising result is that EBs, irrespective of the type of accuracy, produced greater accuracy than the other modes. This result is only partially in accordance with Hypothesis 3 which predicted better differential elevation for EBs over DBs but similar differential accuracy performance for EBs and DBs. Note that we have suggested that differential accuracy assesses within-ratee as well as between-ratee discriminability. If this reasoning is correct, given that the EB mode favors between-ratee discriminability, then the greater differential accuracy of the EB mode over the DB one could be one indication that participants are more attuned to detecting between-ratee

rather than within-ratee differences. Consistent with such a speculation, Beauvois (1987) showed that personality-based evaluations were better predicted by an ordinal model than by either a normalized or dialectical one (see Lamiell, 1981). That is, when people make personality judgments, they simply rank the targets they observe. If this is so, when participants use EBs, they are in a situation of consistency between their intrinsic process and the process implied by the rating format they used (i.e., they focus on ranking ratees and the format helps them do this). In contrast, when they use DBs, they are in a situation of inconsistency between their intrinsic process and the process implied by the format (i.e., they focus on ranking ratees, and the format leads them to focus on differences within ratees). Such consistency should facilitate the detection of between-ratee differences and thus render it more accurate, but it probably does not facilitate the detection of within-ratee differences. Inconsistency between intrinsic processes and rating format would probably impair the detection of both between-ratee and within-ratee differences. As differential accuracy assesses the correct detection of both kinds of differences, EBs should logically produce better DA scores (they improve the detection of between-ratee differences) than DBs (they impair the detection of both kinds of differences) when participants are spontaneously inclined toward the detection of between-ratee distinctions.

If our interpretation is correct, then, if raters are explicitly given a goal which favors the detection of within-ratee differences (a developmental goal) they should improve their differential accuracy by using DBs because there should then be consistency between the process used by the rater and the process implied by the rating format: a focalization on within-ratee differences. To test this post-hoc interpretation, we conducted a second experiment in which we added a manipulation of rating purpose.

Experiment 2

A substantial body of research indicates that administrative and developmental purposes have different effects on a variety of outcomes (see Murphy & Cleveland, 1991), but very few studies have dealt with the effect of these purposes on indexes of discriminability (Wong & Kwong, 2007; Zedeck & Cascio, 1982) or accuracy (Murphy, Kellam, Balzer, & Armstrong, 1984), and their results are rather inconclusive. Thus, our predictions are based on the theoretical argument that the administrative purpose deals mainly with the ranking of ratees and as such should focus the attention of the rater on the differences among ratees, whereas the developmental purpose implies an examination of the different abilities of each ratee and should thus focus the attention of the rater on the differences within a ratee (for a similar argumentation see Murphy & Cleveland, 1991). This reasoning led us to hypothesize that situations of consistency between the purpose and the rating format (administrative with EBs *versus* developmental with DBs; in this study, we omitted the trait format) should increase the discriminability indexes that are related to the purpose (administrative purpose and between-ratee discriminability *versus* developmental purpose and within-ratee discriminability). In consequence, we hypothesized:

Hypothesis 5: An administrative purpose should increase between-ratee discriminability and EB rating scales should be the most efficient mode for this score with this purpose. A developmental purpose should increase within-ratee discriminability, and DB rating scales should be the most efficient mode for this score with this purpose.

In relation to the accuracy scores, we expected:

Hypothesis 6: An administrative purpose, more than a developmental one, should increase DEL scores, and with this purpose, EBs should lead to more accuracy than DBs.

Finally, because DA measures between- and within-ratee discriminability, we hypothesized:

Hypothesis 7: An interaction between the type of purpose and the rating scale format will result such that DA should be higher for EB scales in the presence of an administrative purpose but higher for DB scales when the purpose is developmental.

Finally, in line with Hypotheses 2 and 4 which predicted greater between-ratee discriminability and DEL scores on the agency dimension for the EB rating mode, we expected that this advantage results principally in the administrative purpose condition:

Hypothesis 8: Greater between-ratee discriminability and DEL scores will result on the agency dimension for the EB mode of rating only in the administrative purpose condition.

Method

Participants and Procedure

Eighty students in management in a French university who were familiar with the appraisal interview participated in this experiment in small groups (two to four per group) on a voluntary basis. No incentives were given for participation. All participants viewed the seven videotaped appraisal interviews used in Experiment 1 and rated them on evaluation forms. Two different presentation orders of the videotapes were used to control for possible order effects. Participants were randomly assigned to one of eight different experimental groups obtained from the experimental design of 2 (Purpose of evaluation: Administrative vs. Development) X 2 (Mode of rating: EB scales vs. DB scales) X 2 (Videotape presentation order).

Purpose Manipulation

All participants were informed that the research involved testing a new kind of rating format. They were reminded that appraisal could pursue two main goals: an administrative or

a developmental one. In order to test our new rating modes, they were to focus on only one of these goals. Participants in the administrative condition were told that their "central goal would be to rank people in order to decide whom to promote and whom to warn." Participants in the developmental condition were told that their "central goal would be to rank each manager's competencies in order to determine whether several of these are satisfactory or should be reinforced."

Manipulation Check

A post-experimental questionnaire proposed a series of mutually exclusive true-false questions that indicated participants' understanding of the purpose of the ratings, for the administrative purpose (e.g., "Your ratings focused on the awarding and ranking of managers") and for the development purpose (e.g., "Your ratings focused on the possibilities of development for each manager"). Over 96% of the participants answered the questions appropriately, suggesting that our manipulation of rating purpose was successful.

Results

Preliminary Analyses

As for Experiment 1, we first computed a series of correlations using targets as the unit of analysis. As can be seen (Table 8), the correlations among the ratings for the different dimensions and rating formats were high, showing that each target was rated similarly irrespective of the dimension or the rating format. Again, very high correlations between agency and communion were obtained confirming the potential confound of these two dimensions.

Discriminability Measures

We submitted the within-ratee and the between-ratee discriminability measures to the same 2 (Purpose of rating: *administrative vs. development*) X 2 (Mode of Rating: *EBs vs. DBs*) X 2 (Dimension of scales: *communion vs. agency*) X 2 (Valence of scales: *positive vs.*

negative) ANOVA with the first two variables as between-participants and the latter two as within-participants factors (see Table 9 for the ANOVA summary table).

Within-ratee discriminability. As hypothesized (H 5), the main effect of purpose of rating showed that the developmental purpose produced greater within-ratee discriminability (M = 0.52) than the administrative one (M = 0.27). The main effect of Mode of rating was also significant showing that DBs induced greater within-ratee discriminability (M = .49) than EBs (M = .30). The interaction between Purpose and Mode of rating was significant. It indicated the additive effects of Purpose and Mode of rating. A series of *post-hoc* analyses (Bonferroni test) showed that DB rating scales for the developmental purpose gave rise to higher within-ratee discriminability (M = 0.58) than for all other conditions ($M_{\text{DB administrative}}$ = 0.41; $M_{\text{EB development}}$ = 0.46; $M_{\text{EB administrative}}$ = 0.13), thus confirming hypothesis 5. Finally, a significant interaction between the mode and the dimension of rating emerged but added no supplementary information beyond the mode of rating main effect.

Between-ratee discriminability. As predicted (H 5), a main effect for Purpose of rating resulted, and the administrative purpose permitted higher between-ratee discriminability (M = 1.94) than the developmental one (M = 1.86). The analysis also revealed a main effect of Mode of rating showing that EBs gave rise to greater between-ratee discriminability (M = 1.94) than DBs (M = 1.84). Finally, the main effect of Dimension of rating was also significant showing that between-ratee discriminability was more pronounced for agency ratings (M = 1.92) than for communion ones (M = 1.87). The interaction between rating purpose and dimensions of scales was significant and showed that the main effect of Dimension of rating appeared only in the administrative purpose condition (M_{agency} = 1.96, $M_{\text{communion}}$ = 1.90) but not in the developmental one. Finally, the interaction between purpose of rating, mode of rating and Dimensions of scales was not significant contrary to expectations of Hypothesis 8. However, following the recommendations of Rosnow and Rosenthal (1989)

for testing hypotheses that propose specific comparisons, we computed a set of planned contrasts comparing the discriminability of the EB rating scales in the administrative purpose condition for agency scales to three other conditions: (a) the EB rating scales in the administrative purpose for communion scales, (b) the EB rating scales in the development purpose condition for agency scales, and (c) the DB rating scales in the administrative purpose condition for agency scales. Moreover, as we planned four comparisons, we set the alpha level at $p < .0125$. The planned comparisons showed, as predicted by Hypothesis 8, that the EB rating scales in the administrative purpose condition for the agency scales gave rise to more between-discriminability ($M = 2.03$) than the other conditions ($M_{\text{EB administrative communion}} = 1.93$, $F(1, 76) = 12.45$, $p < .001$; $M_{\text{EB development agency}} = 1.86$, $F(1, 76) = 7.85$, $p < .007$; $M_{\text{DB administrative agency}} = 1.91$, $F(1, 76) = 21.99$, $p < .001$).

As in Experiment 1, we conducted two series (one for each purpose of rating condition) of eight two-way random model effects (variance components), one for each performance dimension X valence X rating format, with raters and targets as random factors and each rating as a dependent variable. We expected to reproduce the results of Experiment 1 in the administrative purpose condition (a higher ratee variance components compared to error and rater variance components for the EB rating format, as well as higher ratee variance components for the EB than for the DB rating format). However, as the developmental purpose condition was incongruent with the processes implied by the EB rating format, we expected the ratee variance component to be lower in this condition. The results indicated, that, for the EB rating format in the administrative purpose condition, the ratee variance components were higher than the error and rater variance components, respectively for positive communion (.95), negative communion (.94), positive agency (.95), and negative agency (.95). Moreover, these were higher than those obtained for the DB rating format (.83, .83, .83, .86). In the developmental purpose condition, the ratee variance components were

generally lower than in the administrative purpose condition and they were not very different between the two rating formats (.64, .59, .62, .72, for EB, and .67, .70, .70, .64 for DB). It seems clear from these results that it is the true score that drives higher between-ratee discriminability for the EB rating format in the administrative purpose condition. Moreover, the EB rating format is more affected by differences among ratees than the DB format. In the developmental purpose condition, the difference between the two rating formats was smaller, and globally, the ratee variance component was lower in this purpose condition than in the administrative one.

Accuracy Measures

The analyses for the two accuracy measures involved the same ANOVA design (Table 9).

Differential elevation. The analysis on DEL revealed a main effect of Purpose of rating showing that the administrative purpose (M = 1.06) gave rise to better accuracy scores than the developmental purpose (M = 1.11). A main effect of Mode of rating showed that EBs (M = 1.06) gave rise to more accurate scores than DBs (M = 1.17). A main effect of Dimension of rating was also significant revealing that Agency scales (M = 1.06) produced more accurate scores than Communion scales *(*M = 1.12). This effect was qualified by an interaction between Dimension and valence of scales showing that this difference was more pronounced on negative rating scales than on positive ones. The expected (Hypothesis 6) interaction between Purpose of rating and Mode of rating was not significant. However, following the recommendations of Rosnow and Rosenthal (1989) for testing hypotheses that propose specific comparisons, we computed a set of planned contrasts comparing the accuracy of the EB rating scales in the administrative purpose condition to every other condition. Moreover, as we planned three comparisons, we chose to set the alpha level at $p <$

.0166. The planned comparisons showed, as predicted by Hypothesis 6, that the score of the EB rating scales in the administrative purpose condition was more accurate (M = 1.03) than the others ($M_{\text{DB administrative}}$ = 1.10, $F(1, 76) = 19.79$, $p < .001$; $M_{\text{EB development}}$ = 1.09, $F(1, 76) = 45.18$, $p < .001$; $M_{\text{DB development}}$ = 1.13, $F(1, 76) = 97.85$, $p < .001$). Finally, the interaction between purpose of rating, mode of rating and Dimensions of scales was not significant, contrary to Hypothesis 8. However, we computed a set of planned contrasts comparing the accuracy of the EB rating scales in the administrative purpose condition for agency scales to three other conditions: (a) the EB rating scales in the administrative purpose for communion scales, (b) the EB rating scales in the development purpose condition for agency scales, and (c) the DB rating scales in the administrative purpose condition for agency scales. As we planned four comparisons, we set the alpha level at $p < .0125$. The planned comparisons showed, as predicted by Hypothesis 8, that the score of the EB rating scales in the administrative purpose condition for agency scales was more accurate (M = 0.99) than the others ($M_{\text{EB administrative communion}}$ = 1.06, $F(1, 76) = 16.02$, $p < .001$; $M_{\text{EB development agency}}$ = 1.05, $F(1, 76) = 6.50$, $p < .0127$; $M_{\text{DB administrative agency}}$ = 1.07, $F(1, 76) = 14.87$, $p < .001$).

Differential accuracy. The analysis on DA revealed a main effect of Purpose of rating showing that ratings were more accurate with the administrative purpose (M = 0.04) than with the developmental one (M = 0.07). The main effect of Mode of rating revealed that DBs (M = 0.05) gave rise to more accurate scores than EBs (M = 0.06). Finally, the interaction between Purpose and Mode of rating was significant showing, as predicted (Hypothesis 7), that with an administrative purpose, EB rating scales gave rise to more accurate scores (M = 0.03) than DB rating scales (M = 0.06); $F(1, 76) = 34.84$, $p < .001$, $\eta^2 = 0.33$. Inversely, with a developmental purpose, it was the DB rating scales that gave rise to more accurate scores (M = 0.05) than those obtained with the EB rating scales (M = 0.09); $F(1, 76) = 99.09$, $p < .001$, $\eta^2 = 0.60$. Finally, EBs gave rise to more accurate scores with an administrative purpose (M

= 0.03) than with a developmental one (M = 0.09), $F(1, 76) = 332.01$ $p < .001$, $\eta^2 = 0.85$, whereas DBs gave rise to more accurate scores with a developmental purpose (M = 0.05) than with an administrative one (M = 0.06), $F(1, 76) = 15.67$, $p < .001$, $\eta^2 = 0.09$.

Discussion

The results of this study confirmed our hypotheses. As expected, manipulating the purpose of evaluation reinforced the effects of the rating modes when these modes were consistent with the purpose: A developmental purpose enhanced the within-ratee discriminability of DBs whereas an administrative purpose strengthened the between-ratee discriminability of EBs. In contrast, when the purpose of evaluation was inconsistent with the mode used (development purpose with EBs and administrative purpose with DBs), the discriminative power of the mode was inhibited (EBs were less between-ratee discriminant and DBs were less within-ratee discriminant).

As in Experiment 1, between-ratee discriminability was enhanced when evaluations were made on agency dimensions. These results converge with those obtained in the literature (Funder et al., 1987; Mignon et al., 2002; Paulhus et al., 1995). Once again, this result is noteworthy considering the fact that the strong correlations obtained between these two dimensions could have obscured the effect of agency on between-ratee discriminability.

Contrary to Experiment 1, we did not observe any effects for the valence of rating scales. Thus, one possibility is that the initial effect resulted by chance. However, given this uncertainty, it seems necessary to test the valence of the rating dimensions in future research on the discriminability of EBs and DBs.

Finally, the main goal of this second experiment focused on rating accuracy. As expected, the results of Experiment 1 (the systematic superiority of EBs irrespective of the kind of accuracy considered) could be explained by the fact that in that experiment, the participants spontaneously activated an administrative purpose for their evaluation. When

controlling for the purpose, the results on the accuracy scores were in greater accordance with our expectations. DEL seems to be a kind of accuracy which is linked with an administrative purpose (main effect of purpose of evaluation) and with the rating scale mode and the dimension of evaluation which favor this purpose: EBs and agency. However, it should be noted as a limitation that although the planned contrasts implied by Hypotheses 6 and 8 were significant and went in the expected direction, they were embedded in non-significant interactions. The pattern is different for DA, which seemed to pertain to both administrative and developmental purposes. Indeed, when an administrative purpose was activated, it was the EBs which led to more accuracy than DBs, whereas the opposite resulted when a developmental purpose was activated. Although not in contradiction with our hypotheses, the main effects of purpose and mode of rating were not expected. However, these effects are compatible with our assumption that ranking is the default mode of person perception (Beauvois, 1987). Indeed, when the purpose is administrative and the mode of rating is EB, the situation of evaluation is consistent with people's normal functioning; this consistency creates a condition which enables raters to rate more accurately.

General Discussion and Conclusion

The main goal of these experiments was to establish a link between purpose of evaluation, rating scale mode or format, and the psychological processes they activate while examining the accuracy of the ratings resulting from these processes. Globally, the results support the hypothesis that the two fundamental kinds of evaluation purpose (administrative and developmental) produce two distinct effects, these effects being facilitated by the kind of rating mode used.

First, an administrative purpose produced a focus on the differences between the ratees. This focus is consistent with an administrative goal because in order to determine which employee or candidate to punish (reject, propose a disciplinary action) or promote

(promotion decision, merit increase), one has to distinguish among ratees which one is better or worse than others. It is important to note that this purpose forms the basis of what it means to evaluate. In consequence, using a rating mode based on statements (EBs) that were conceived expressly for communicating the value of people facilitates reporting between-ratee differences. One interesting result of the present experiments is that the agency dimension of evaluation seems particularly suited for this kind of goal. Interestingly, the results of the present studies suggest that the administrative purpose corresponds to the default option when people evaluate others.

Second, the developmental purpose seemed to produce a focus on differences within ratees. With this purpose, it is necessary for the rater to assess the strengths and weaknesses of each ratee, so the attention of the rater should be diverted from the differences between people to focus on the within-ratee differences. Such a process is close to descriptive knowledge; that is, reporting the objective characteristics of an object. In consequence, the use of a rating mode based on statements (DBs) that focus on the descriptive characteristics of people should facilitate reporting within-ratee differences. Although the rating forms were presented to the participants before they viewed the videos, we gathered no direct evidence to permit concluding that it was the very fact of consulting the scales that led participants to differential observation, encoding, or storing of performance information. The rating format may also have had its impact at retrieval, helping the participants to structure their impressions in a way congruent with the type of knowledge activated (evaluative or descriptive). Future research could focus on distinguishing between alternative explanations of where in the cognitive processes of rating the mode or format intervenes.

It is important to note that the increase in discriminability produced with both rating purposes and both rating modes should not be interpreted as a biased treatment of information. On the contrary, each purpose and rating mode was better than the other on some

form of accuracy. Because EBs and the administrative purpose focus raters on differences between ratees, they lead to greater differential elevation, the kind of accuracy which deals with discrimination among ratees. The pattern of results is different for differential accuracy which relates to discrimination both among and within rates. Here, rating mode interacted with the type of purpose such that EBs led to better DA scores than DBs when an administrative purpose was activated, whereas DBs led to better DA scores than EBs when a developmental purpose was activated. It should be noted that the results related to the accuracy of EBs contradict the criticism made toward the pragmatic views of social perception made, for example, by Funder (1987): "Accuracy is not viewed as dependent on any properties that the target of judgment actually has. Nearly all the focus is on the judge, not the judged" (p. 656). Respecting this criticism, one could argue that it would be unfair to make administrative personnel decisions based on EBs because this format can be viewed as only reflecting individual raters' idiosyncratic opinions. The results on the accuracy measures as well as those obtained in the variance components analyses (showing that between-ratee discriminability obtained with EBs reflected more true between variance than error variance than that obtained with either DBs or traits) contradict this view by showing that social affordances (evaluations made on EBs) truly capture the properties of the object and are not in the eyes of the beholder (Zebrowitz & Collins, 1997).

Finally, the results showed that agency interacted with the rating format (more particularly on EB rating scales) to increase accuracy (on the DEL component). However, it should be noted that this effect was obtained in both studies even in the presence of strong correlations between communion and agency. Such correlations are not especially new in this literature. They are often interpreted by the fact that agency and communion are highly evaluative constructs (Suitner & Maass, 2008) and that, as a consequence, valence represents a "third" variable that potentially confounds the relation between the two dimensions. In our

experiments, the trait and the EB scales were highly evaluative. This characteristic, in conjunction with the fact that the task that participants performed was an evaluation, probably heightened the correlations between agency and communion. This interpretation is supported by the fact that in both studies the correlations between agency and communion for traits and EBs were always higher than those obtained for DBs which constituted more descriptive scales. Nonetheless, the usefulness of these two dimensions for performance appraisal may be questioned given the presence of such strong correlations between the constructs. It would be interesting to use more distinctive dimensions in the future. For example, researchers (Funder & Dobroth, 1987; Paulhus & Reynolds, 1995) have shown that extraversion, as agency, yielded high target variance (in our terms between-ratee discriminability), this dimension, in conjunction with conscientiousness (two important dimensions in performance appraisal) could be used to explore our hypotheses.

Our results have very interesting implications for performance appraisal and other individual assessment practices. However, any conclusions drawn from the current studies must be limited by the fact that these experiments were only simulations. Supplementary tests should be considered before these rating formats can be used in the field. First of all, the EBs and DBs used were not derived from a job analysis conducted on a real job. Nonetheless, many of the communion behaviors correspond to job performance dimensions relating to team work, personal discipline, or counterproductive work behaviors that are included in various taxonomies of job performance (see Borman et al., 2010 for an integrative review). The agency behaviors are also akin to such performance dimensions as technical proficiency in some jobs and personal discipline, for example. More research should examine the extent to which the nature of the performance dimension, such as organizational citizenship behaviors, job proficiency, or counterproductive work behaviors (cf. Borman et al., 2010) influence the accuracy of the EB or DB ratings. Furthermore, when developing behaviorally

based instruments for performance appraisal, such as the BOS or BARS approaches, it would be possible to develop both EB and DB items during the phase of generating the behavioral exemplars for the identified performance dimensions. To produce them, one could ask people to generate behavioral exemplars from identified performance with inductive sentences such as, for DBs, "someone who…" and, for EB or "someone for whom…." This distinction could therefore contribute to making these rating formats more adapted to various administrative or developmental rating purposes, be they for use in selection interview, performance feedback, individual assessment, or other appraisal contexts.

A second set of limitations concerns the rating context. Although a number of steps were taken in the design of these studies to increase their external validity, the fact remains that the situational influences on performance ratings that are present in organizational settings were not likely to be active in the present experiments. Most notably, the raters did not know the ratees, as is the case in any organization. Thus, they are not accountable (Curtis, Harvey, & Auden, 2005) for their ratings. Also, they did not have the opportunity to observe the ratees during their work. Finally, they made their ratings immediately after the interview, thus processing in an on-line manner, while it has been recognized that performance appraisal is typically a memory-based rating process. Nonetheless, the experimental methodology used here allowed for careful control of the rating stimuli and for calculating accuracy measures, which is less feasible in field research.

Using the videotape of an appraisal interview as stimuli was also problematic. Although we excluded a maximum of interactional characteristics from the videotapes in order to present a majority of information on performance, it remains that, in this situation, the rater did not evaluate the behaviors and the performances of the ratee but her/his intentions or explanations. This constitutes a fundamental difference with the situation of performance appraisal and a potential limitation of the present research. However, it should be

noted that the present results reproduced those of Mignon and Mollaret (2002) which were obtained in a very different situation (a zero acquaintance paradigm). This replication is an argument in favor of the fact that what we can learn from the present research is not restricted to the appraisal interview but could be extended to more general situations of evaluation. Nonetheless, the specific delimitation of the validity of the present results requires further research.

To conclude, it must be stressed that the hypotheses formulated in this paper dealt mostly with the psychometric properties and accuracy of performance ratings. However, it has been shown that ratees' and raters' reactions to an appraisal system might yield a more significant contribution to sustaining the viability of an appraisal system than its psychometric qualities (Cawley, Keeping & Levy, 1998). So, future work should also deal with the satisfaction raters and ratees might feel toward the EB and DB rating modes, perhaps by focusing on the distributive and procedural justice (cf. Greenberg, 1986) of these different formats as a function of their purpose (Roch, Sternburg, & Caputo, 2007).

References

Abele, A.E., Uchronski, M., Suitner, C., & Wojciszke, B. (2008). Toward an operationalization of the fundamental dimensions of agency and communion: trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology, 38,* 1202-1217. Doi: 10.1002/ejsp.575

Bassili, J. N. (1989). Traits as action categories versus traits as person attributes in social cognition. In J. N. Bassili (Ed.), *On line cognition in person perception*. Hillsdale, NJ: Lawrence Erlbaum.

Beauvois, J. L. (1987). The intuitive personologist and the individual differences model. *European Journal of Social Psychology, 17,* 81-94. Doi:10.1002/ejsp.2420170108

Beauvois, J. L. & Dubois, N, (2000). Affordances in social judgment: Experimental proof of why it is a mistake to ignore how others behave towards a target and look solely at how the target behaves. *Swiss Journal of Psychology, 59,* 16-33. Doi:10.1024//1421-0185.59.1.16

Bernardin, H., & Beatty, R.W. (1984). *Performance appraisal: Assessing human behavior at work*. Boston: Kent.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior & Human Performance, 20,* 238-252. Doi:10.1016/0030-5073(77)90004-6

Borman, W. C., Bryant, R. H., & Dorio, J. (2010). The measurement of task performance as criteria in selection research. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 439-461). New York: Routledge.

Cambon, L., Djouari, A., & Beauvois, J. L. (2006). Social judgment norms and social utility: When it is more valuable to be useful than desirable. *Swiss Journal of Psychology, 65*, 167-180. Doi: 10.1024/1421-0185.65.3.167

Cascio, W. F., & Aguinis, H. (2008). Research in industrial and organizational psychology from 1963 to 2007: Changes, choices, and trends. *Journal of Applied Psychology, 93*, 1062-1081. Doi: 10.1037/0021-9010.93.5.1062

Cawley, B. D., Keeping, L. M., & Levy, P. E. (1998). Participation in the performance appraisal process and employee reactions: A meta-analytic review of field investigations. *Journal of Applied Psychology, 83*, 615-633. Doi: 10.1037//0021-9010.83.4.615

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. Doi: 10.1037//0021-9010.74.1.130

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin, 52*, 177-193. Doi: 10.1037/h0044919

Curtis, A. B., Harvey, R. D., & Ravden, D. (2005). Sources of political distorsions in performance appraisals: Appraisal purpose and rater accountability. *Group and Organization Management, 30*, 42-60. Doi: 10.1177/1059601104267666

DeNisi, A. S. & Peters, L. H. (1996). Organization of information in memory and the performance appraisal process: Evidence from the field. *Journal of Applied Psychology, 81*, 717-737. Doi: 10.1037//0021-9010.81.6.717

DeNisi, A.S., Robbins, T.L., & Summers, T.P. (1997). Organization, processing, and use of performance information: A cognitive role for appraisal instruments. *Journal of Applied Social Psychology, 27,* 1884-1905. Doi: 10.1111/j.1559-1816.1997.tb01630.x

DeYoung, C.G., Quilty, L.C., Peterson, J.B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93,* 880-896. Doi: 10.1037/0022-3514.93.5.880

Dubois, N., & Tarquinio, C. (1998). Le traitement de l'information évaluative par des professionnels de l'évaluation sociale. (Evaluative information processing by professionnals of social evaluation). *Revue Internationale de Psychologie Sociale, 11,* 99-122.

Fay, C. H., & Latham, G. P. (1982). Effects of training and rating scales on rating errors. *Personnel Psychology, 35,* 105-116. Doi: 10.1111/j.1744-6570.1982.tb02188.x

Funder, D.C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101,* 75-90. Doi: 10.1037//0033-2909.101.1.75

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology, 52,* 409-418. Doi: 10.1037//0022-3514.52.2.409

Goffin, R.D., & Olson, J.M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science, 6,* 48-60. Doi: 10.1177/1745691610393521

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26-34. Doi: 10.1037//0003-066X.48.1.26

Greenberg, J. (1986). Determinants of perceived fairness of performance evaluations. *Journal of Applied Psychology, 71,* 340-342. Doi: 10.1037//0021-9010.71.2.340

Heneman, R. L. (1988). Traits, behaviors, and rater training: Some unexpected results. *Human Performance, 1,* 85-98. Doi: 10.1207/s15327043hup0102_1

John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66-100). New York: Guilford.

Judd, C., James-Hawkins, L. Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology, 89*, 899-913. Doi: 10.1037/0022-3514.89.6.899

Lamiell, J. T. (1981). Toward an idiothetic psychology of personality. *American Psychologist, 36*, 276-289. Doi: 10.1037//0003-066X.36.3.276

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72-107. Doi: 10.1037//0033-2909.87.1.72

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. New York: Academic Press.

Latham, G.P., & Wexley, K.N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30,* 255-268. Doi: 10.1111/j.1744-6570.1977.tb02092.x

McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review, 90*, 215-238. Doi: 10.1037//0033-295X.90.3.215

McDonald, T. (1991). The effect of dimension content on observation and ratings of job performance. *Organizational Behavior and Human Decision Processes, 48*, 252-271. Doi: 10.1016/0749-5978(91)90014-K

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46. Doi: 10.1037//1082-989X.1.1.30

Mignon, A., & Mollaret, P. (2002). Applying the affordance conception of traits: A person perception study. *Personality and Social Psychology Bulletin, 28*, 1327-1334. Doi: 10.1177/014616702236825

Montoya, R.M., & Horton, R.S. (2004). On the importance of cognitive evaluation as a determinant of interpersonal attraction. *Journal of Personality and Social Psychology, 86*, 696-712. Doi: 10.1037/0022-3514.86.5.696

Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. *Journal of Applied Psychology, 71*, 39-44. Doi: 10.1037//0021-9010.71.1.39

Murphy, K. R., & Constans, J. I. (1987). Behavioral anchors as a source of bias in rating. *Journal of Applied Psychology, 72*, 573-577. Doi: 10.1037//0021-9010.72.4.573

Murphy, K. R., & Pardaffy, V. A. (1989). Bias in Behaviorally Anchored Rating Scales: Global or scale-specific? *Journal of Applied Psychology*, *74*, 343-346. Doi: 10.1037//0021-9010.74.2.343

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

Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? *Journal of Applied Psychology*, *67*, 562-567. Doi: 10.1037//0021-9010.67.5.562

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology, 67*, 320-325. Doi: 10.1037//0021-9010.67.3.320

Murphy, K. R., Kellam, K. L., Balzer, W. K., & Armstrong, J. G. (1984). Effects of the purpose of rating on accuracy in observing teacher behavior and evaluating teaching

performance. *Journal of Educational Psychology, 76*, 45-54. Doi: 10.1037//0022-0663.76.1.45

Paulhus, D. L., & Reynolds, S. (1995). Enhancing target variance in personality impressions: Highlighting the person in person perception. *Journal of Personality and Social Psychology, 69*, 1233-1242. Doi: 10.1037//0022-3514.69.6.1233

Piotrowski, M. J., Barnes-Farrell, J. L., & Esrig, F. H. (1989). Behaviorally anchored bias: A replication and extension of Murphy and Constans. *Journal of Applied Psychology, 74*, 823-826. Doi: 10.1037//0021-9010.74.5.823

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69*, 581-588. Doi: 10.1037//0021-9010.69.4.581

Pulakos, E. D., Schmitt, N. & Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within ratees to measure halo. *Journal of Applied Psychology, 71*, 29-32. Doi: 10.1037//0021-9010.71.1.29

Roch, S. G., Sternburg, A. M., & Caputo, P. M. (2007). Absolute vs. relative performance rating formats: Implications for fairness and organizational justice. *International Journal of Selection and Assessment, 15*, 302-316. Doi: 10.1111/j.1468-2389.2007.00390.x

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological sciences. *American Psychologist, 44*, 1276-1284. Doi: 10.1037//0003-066X.44.10.1276

Singh, R., Simons, J.P, Young, D.P., Sim, B.S., Chai, X.T., Singh, S., & Chiou, S.Y. (2009). Trust and respect as mediators of the other- and self-profitable trait effects on interpersonal attraction. *European Journal of Social Psychology, 39*, 1021-1038. Doi: 10.1002/ejsp.605

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149-155. Doi: 10.1037/h0047060

Srull, T. K., & Wyer, R. S., Jr. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology, 37*, 1660-1672. Doi: 10.1037//0022-3514.37.10.1660

Suitner, C., Maass, A. (2008). The role of valence in the perception of agency and communion. *European Journal of Social Psychology, 38*, 1073-1082. Doi: 10.1002/ejsp.525

Tziner, A., & Kopelman, R. E. (2002). Is there a preferred performance rating format? A non-psychometric perspective. *Applied Psychology: An International Review, 51*, 479-503. Doi: 10.1111/1464-0597.00104

Woehr, D. J., & Roch, S. (2012). Supervisory performance ratings. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 517-531). Oxford: Oxford University Press. Doi: 10.1093/oxfordhb/9780199732579.013.0022

Wojciszke, B., Abele, A. E., & Baryla, W. (2009). Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency. *European Journal of Social Psychology, 39*, 973-990. Doi: 10.1002/ejsp.595

Wong, K. F. E., & Kwong, J. Y. Y. (2007). Effects of rater goals on rating patterns: Evidence from an experimental field study. *Journal of Applied Psychology, 92*, 577-585. Doi: 10.1037/0021-9010.92.2.577

Zebrowitz, L. A., & Collins, M. A. (1997). Accurate social perception at zero acquaintance: The affordances of a Gibsonian approach. *Personality and Social Psychology Review, 1*, 204-223. Doi: 10.1207/s15327957pspr0103_2

Zedeck, S., & Cascio, W.F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology, 67,* 752-758. Doi: 10.1037//0021-9010.67.6.752

Footnotes

[1] A third important purpose is a research function, notably when researchers are conducting criterion-related validation studies. Although our experiments do not deal with this kind of purpose, it is important to note that the concern in criterion-related validation studies is between-ratee discriminability and not within-discriminability. Thus, the results of the present research could also have implications for various research purposes for appraisal.

[2]Murphy and Balzer (1986) obtained results that contradict this hypothesis. They showed that ratings had greater DEL with specific (behavior based format) rather than global rating items (trait based format). However, as their global and specific items were not comparable (global items were not derived from their specific items), we did not take these results into account to formulate our hypotheses.

[3] The behaviors submitted to the pretests (N_{EB} =60, N_{DB} =60) were extracted from the most frequent behaviors in a large pool of EBs and DBs obtained in a pilot study in which 80 undergraduate students indicated the first EB (versus DB) that came to mind for the set of 40 traits initially selected.

[4] These pilot tests were conducted with two groups of graduate students who were blind to the intended performance levels of each tape. The first group (N = 11) ranked the 7 managers according to their performance level. The second group (N = 14) ranked the 12 scales for each manager from the one on which the manager was the most competent to the one on which the manager was the least competent. For the results of the first test (ranking of the managers), although all differences were not significant, the managers were ranked as intended. For the results of the second test (ranking of the scales for each manager), although there was some variability in the ranking of the scales, in every case, a scale on which a manager's performance was effective always received a higher ranking than a scale on which a manager's performance was ineffective. It should be noted that although participants

perceived differences in performance on the scales when the differences were great (i.e., differences over 4 ranks), they did not identify more subtle differences in performance (i.e., differences of 1 and 2 ranks). But this was not problematic. The important result was that the differences between the scales were sufficiently explicit so as to allow the participants to achieve discriminability within the targets.

[5]The order of presentation of the 7 targets was excluded from all the analyses in studies 1 and 2 because no main or interaction effect implying it was significant.

[6]Using restricted maximum likelihood estimation did not change the results in study 1 or in study 2.

[7]We are grateful to an anonymous reviewer to have suggested this test.

Table 1

English Translations of French (in Parentheses) for some example items of the Three Rating Modes

Type	Trait	Descriptive Behaviors (DBs)	Evaluative Behaviors (EBs)
C+	Sociable (sociable)	Someone who likes to work with others (quelqu'un qui aime travailler avec les autres)	Someone with whom it is easy to work (quelqu'un avec qui il est facile de travailler)
C-	Aggressive (agressif)	Someone who talks nastily (quelqu'un qui parle méchamment)	Someone with whom it is easy to quarrel (quelqu'un avec qui il est facile de se disputer)
A+	Conscientious (consciencieux)	Someone who checks that his/her work does not include errors (quelqu'un qui vérifie que son travail ne comprenne pas d'erreurs)	Someone to whom you could entrust the proofreading of your work (quelqu'un en qui on peut avoir confiance pour reviser un travail)
A-	Irresponsible (irresponsable)	Someone who doesn't attend to important things to do (quelqu'un qui ne prète pas attention aux choses importantes à faire)	Someone on whom you can't rely (quelqu'un à qui on ne peut pas faire confiance)

NOTE: Types of traits: A = Agency, C = Communion, - = Negative, + = Positive.

Table 2

Single-rater (ICC(C,1)) intraclass correlations for expert ratings (judge) and for participants ratings (subject) as a function of Mode of rating format, dimensions of scales and valence of scales for study 1.

	EB rating format				DB rating format				Traits rating format			
	Communion		Agency		Communion		Agency		Communion		Agency	
	+	-	+	-	+	-	+	-	+	-	+	-
Judge ICC(C,1)	.92	.90	.92	.92	.68	.69	.68	.69	.88	.69	.68	.69
Subject ICC(C,1)	.54	.66	.72	.67	.36	.31	.34	.39	.40	.29	.51	.40

NB: + for positive scales, - for negative scales

Table 3

Single-rater (ICC(C,1)) intraclass correlations for expert ratings (judge) and for participants ratings (subject) as a function of Mode of rating format, dimensions of scales, valence of scales, and purpose condition in study 2.

		EB rating format				DB rating format			
		Communion		Agency		Communion		Agency	
		positive	negative	positive	negative	positive	negative	positive	negative
Administrative	Judge ICC(C,1)	.95	.95	.94	.94	.88	.92	.89	.91
purpose	Subject ICC(C,1)	.82	.85	.82	.81	.72	.75	.76	.67
Development	Judge ICC(C,1)	.94	.94	.94	.94	.91	.89	.93	.89
purpose	Subject ICC(C,1)	.67	.67	.68	.69	.54	.51	.53	.65

Table 4

Means, Standard Deviations, Reliabilities, and Inter-correlations among the Rating Formats and the Dimensions of Evaluation (N = 7)

Rating Format	Dimension	Mean	Std. Dev.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Trait	C+ (1)	4.42	1.22	(.92)											
	C- (2)	4.61	1.19	.81**	(.89)										
	A+ (3)	4.41	1.31	.67*	.69*	(.94)									
	A- (4)	4.62	1.39	.74*	.71*	.89**	(.92)								
EB	C+ (5)	3.78	1.66	.80**	.69*	.87**	.90**	(.94)							
	C- (6)	4.03	1.91	.72*	.72*	.88**	.88**	.96**	(.96)						
	A+ (7)	3.95	1.98	.74*	.69*	.92**	.87**	.89**	.89**	(.96)					
	A- (8)	3.98	1.90	.74*	.74*	.92**	.90**	.93**	.91**	.96**	(.95)				
DB	C+ (9)	3.33	1.21	.82**	.42	.60	.67*	.71*	.56	.60	.63	(.91)			
	C- (10)	4.81	1.31	.32	.92**	.67*	.72*	.64	.67*	.62	.70*	.66*	(.89)		
	A+ (11)	4.04	1.56	.61	.65	.73*	.71*	.69*	.72*	.72*	.80**	.39	.66*	(.94)	
	A- (12)	4.43	1.39	.62	.77*	.87**	.90**	.78*	.80**	.80**	.87**	.60	.49	.80**	(.92)

Note. A+ = positive agency, A- = negative agency, C+ = positive communion, C- = negative communion, EB = Evaluative Behaviors rating format, DB = Descriptive Behaviors rating format. Values in parentheses indicate the reliability score for the scale. * $p < .05$, ** $p < .01$

Table 5

ANOVA Summary Table (experiment 1)

	Discriminability measures					Accuracy Measures				
	Within-ratee		Between-ratee			Differential Elevation			Differential Accuracy	
Type of effects	discriminability		discriminability							
	$F(2, 57)$	η^2	$F(2, 57)$	$F(1, 57)$	η^2	$F(2, 57)$	$F(1, 57)$	η^2	$F(2, 57)$	η^2
Mode of rating (1)	30.92**	.52	11.83**		.30	116.82**		.80	45.52**	.62
Dimension of scales (2)	1.81	.02		5.46*	.08		95.70**	.54	1.65	.01
Valence of scales (3)	1.60	.02		4.31*	.07	1.21		.01	1.043	.01
(1) X (2)	2.03	.06	3.04		.08	12.17**		.13	7.15**	.28
(1) X (3)	0.32	.01	1.27		.03	1.14		.03	1.49	.02
(2) X (3)	0.72	.01	0.12		.001	0.27		.003	0.35	.005
(1) X (2) X (3)	2.17	.06	5.17**		.15	7.38		.21	0.28	.01

*$p < .05$, ** $p < .01$.

Table 6

Mean Between-ratee Discriminability Scores as a Function of Mode of Rating, Dimension of Scales, and Valence of Scales (Experiment 1)

	Dimension of scales			
	C+	C-	A+	A-
Trait	1.85_{a}	1.83_{a}	1.74_{a}	1.97_{a}
EB	2.10_{b}	2.15_{b}	2.37_{d}	2.32_{c}
DB	1.85_{a}	2.06_{b}	1.99_{a}	1.99_{a}

Note. A+ = positive agency, A- = negative agency, C+ = positive communion, C- = negative communion, EB = Evaluative Behaviors rating format, DB = Descriptive Behaviors rating format. For each comparison, means with different subscripts are significantly different at $p < .05$.

Table 7

Mean Differential Elevation Scores as a Function of Mode of Rating, Dimension of Scales, and Valence of Scales (Experiment 1)

	Dimension of Scale			
	C+	C-	A+	A-
Trait	0.44_{e}	0.52_{f}	0.37_{d}	0.31_{cd}
EB	0.21_{b}	0.12_{a}	0.12_{a}	0.13_{a}
DB	0.51_{ef}	0.46_{ef}	0.29_{c}	0.33_{cd}

Note. A+ = positive agency, A- = negative agency, C+ = positive communion, C- = negative communion, EB = Evaluative Behaviors rating format, DB = Descriptive Behaviors rating format. For each comparison, means with different subscripts are significantly different at $p < .05$.

Table 8 *Means, Standard Deviations, Reliabilities, and Inter-correlations among the Rating Formats and the Dimensions of Evaluation for the Administrative and Development Purpose (N = 7)*

Type of purpose	Rating format	Dimension	Mean	Std. Dev.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		A+ (1)	3.92	1.94	(.94)							
		A- (2)	3.93	2.00	.94**	(.93)						
	EB	C+ (3)	3.93	1.90	.89**	.88**	(.94)					
Administrative		C- (4)	3.94	1.91	.87**	.86**	.93**	(.96)				
purpose		A+ (5)	4.16	1.80	.79**	.81**	.86**	.78*	(.95)			
		A- (6)	4.01	1.81	.81**	.83**	.84**	.81**	.89**	(.92)		
	DB	C+ (7)	4.09	1.75	.80**	.83**	.81**	.84**	.79**	.79**	(.93)	
		C- (8)	4.01	1.66	.78*	.83**	.82**	.85**	.66*	.81**	.90**	(.94)
		A+ (1)	3.60	1.43	(.93)							
		A- (2)	3.90	1.71	.92**	(.93)						
	EB	C+ (3)	3.72	1.50	.81**	.84**	(.93)					
Development		C- (4)	3.75	1.43	.81**	.85**	.92**	(.93)				
purpose		A+ (5)	3.91	1.62	.81**	.81**	.79**	.75*	(.94)			

	A- (6)	3.92	1.56	.80**	.82**	.81**	.79**	.87**	(.96)		
DB	C+ (7)	3.90	1.60	.79**	.83**	.83**	.82**	.67*	.61	(.94)	
	C- (8)	3.89	1.64	.75*	.83**	.83**	.84**	.58	.73*	.87**	(.94)

Note. A+ = positive agency, A- = negative agency, C+ = positive communion, C- = negative communion, EB = Evaluative Behaviors rating format, DB = Descriptive Behaviors rating format. Values in parentheses indicate the reliability score for the scale. * $p < .05$, ** $p < .01$

Table 9

ANOVA Summary Table (experiment 2)

	Discriminability measures				Accuracy Measures			
Type of effects	Within-ratee discriminability		Between-ratee discriminability		Differential Elevation		Differential Accuracy	
	$F(1, 76)$	η^2	$F(1, 76)$	η^2	$F(1, 76)$	η^2	$F(1, 76)$	η^2
Purpose of rating (1)	380.61**	.52	28.08**	.19	27.34**	.18	157.57**	.28
Mode of rating (2)	227.45**	.31	32.64**	.23	38.69**	.27	14.24**	.02
Dimension of scales (3)	0.52	.005	4.30*	.04	62.40**	.45	1.69	.04
Valence of scales (4)	1.33	.01	0.60	.005	1.65	.01	0.06	.001
(1) X (2)	36.51**	.05	5.60*	.04	2.80	.01	333.86**	0.64
(1) X (3)	0.66	.007	5.79*	.06	0.79	.004	0.41	.009
(1) X (4)	1.11	.01	2.04	.02	1.13	.01	2.52	.06
(2) X (3)	4.88*	.05	2.15	.02	1.93	.01	0.44	.01
(2) X (4)	0.002	.00001	0.12	.001	1.26	.01	0.07	.001

(3) X (4)	2.50	.02	1.71	.02	13.39**	.14	0.61	.01
(1) X (2) X (3)	0.67	.007	0.01	.00001	0.15	.001	0.34	.008
(1) X (2) X (4)	1.50	.01	2.54	.03	1.37	.01	0.26	.006
(1) X (3) X (4)	0.02	.00001	0.0001	.00001	3.50	.03	0.05	.001
(2) X (3) X (4)	0.3	.002	0.27	.003	0.92	.009	0.38	.008
(1) X (2) X (3) X (4)	2.38	.02	0.92	.02	0.001	.00001	0.002	.00001

*$p < .05$, ** $p < .01$.