

Rating Accuracy and Performance

Darrin Kass^a

Jung Seek Kim^b

Shiloh Erdley^c

Steven Welch^e

Bloomsburg University
Bloomsburg, PA 17821
USA

Paul F. Rotenberry^d

^dWestchester University
700 South High Street, West Chester, PA 19383
USA

Abstract

Purpose: This paper aims to explore the relationship between self-other rating agreement (SOA) and leadership and academic performance. **Design:** A sample of 129 students enrolled in a part-time, regional MBA program were divided into three SOA categories: Self-aware, Underraters, and Overraters. Performance was evaluated with the completion of a managerial in-basket assessment and final MBA GPA. **Findings:** Leadership and academic performance varied by SOA, with underraters exhibiting the highest scores, followed by self-aware, and then overraters. A key finding was that underraters, not self-aware raters, demonstrated the highest levels of performance, raising questions about the use of accurate self-assessments as an indicator of leadership effectiveness. **Conclusion:** Practical and research implications are discussed for the mounting evidence that underrating may be more important for leadership than self-awareness. Identifying and developing high potential leaders is a central concern of organizations, and therefore, organizations should identify underraters in the selection and promotion processes.

Keywords: rating accuracy, underrating, performance

Rating Accuracy and Performance: We Like Leaders that Underrate Themselves, but are they Better Performers?

1. Introduction

Research has consistently found that self-other rating agreement (SOA) has important implications for leadership outcomes. The most positive leadership outcomes are associated with self-aware individuals, those whose self-ratings match those of their observers (e.g., Yammarino & Atwater, 1997), compared to those who over or underestimate their leadership ability. The consistency of these findings, along with the general belief about the value of self-awareness for leadership effectiveness, has led some to conclude that, “self-awareness is the most important skill to be successful in the 21st century at work” (Eurich, 2018, p. 1). More recently, however, research has shown that underrating, although technically considered a rating error, is also associated with high levels of leadership effectiveness, and that these ratings may be even higher than self-aware individuals. This study sought to explore further the impact of SOA on leadership outcomes using more objective measures of both SOA and leadership performance. This eliminates many of the biases associated with past research and allows for a more precise investigation of SOA on leadership outcomes.

2. Background

Research suggests that self-ratings of ability, while commonly utilized, are laden with inaccuracies (Atwater & Yammarino, 1997; Fleener, Smither, Atwater, Braddy & Sturm, 2010; Harris & Schaubroeck, 1988; Mabe & West, 1982; Moshavi, Brown & Dodd, 2003; Zenger & Folkman, 2015). Generally speaking, self-ratings lack accuracy, reliability, and validity (Ashford, 1989; Harris & Schaubroeck, 1988; Mabe & West, 1982). In general, how we see ourselves does not match how other people see us, which occurs for various internal and external factors (see Fleener et al., 2010, for a review). Zenger (2014) suggested that self-ratings only match external ratings 50 percent of the time, even when objective measures are utilized.

In the context of rating accuracy, the congruence between self- and other ratings is viewed as an indicator of self-awareness (Fleener et al., 2010). Therefore, self-aware individuals see themselves as others see them. However, not all individuals are self-aware, and they vary in the degree to which they under or overestimate their abilities. Individuals who underestimate their skills and evaluate themselves with lower scores than observers are categorized as underraters, while those who overestimate their abilities and evaluate themselves with higher scores are classified as overraters.

There are important differences in the leadership outcomes associated with each of these different categorizations of SOA. In a review of the literature, Yammarino and Atwater (1993) noted that, “accurate self-perceivers have been found to be more successful (regardless of how success was defined) than either over-estimators or under-estimators...” (p. 240). Accurate raters were more effective, made better decisions, were more likely to be promoted, and exhibited the highest performance levels (Yammarino & Atwater, 1997). The relationship between rating accuracy and positive outcomes was so strong that the authors recommended that, “high self-other rating agreement should be a goal of all training programs that use feedback as a development tool” (p. 43). The concept of self-awareness has gained a significant amount of attention in the leadership literature, particularly since leadership effectiveness has been associated with self-awareness (Bass & Yammarino, 1991). The authentic leadership approach is centrally grounded in leader self-awareness, and George, Sims, McLean, and Mayer (2007) reported the following: “When the 75 members of the Stanford Graduate School of Business’s Advisory Council were asked to recommend the most important capability for leaders to develop, the answer was nearly unanimous: self-awareness” (pg. 133).

Overrating has consistently been associated with the most negative individual and organizational outcomes. Overraters tend to have more negative attitudes, lack awareness of their strengths and weaknesses, and are less likely to utilize training and development opportunities (Woo, Sims, Rupp, & Gibbons, 2008). Overrating is also linked to poor organizational effectiveness, career derailment, and negative physical health (Goffin & Anderson, 2002; Vecchio & Anderson, 2009; Woo et al., 2008). Overrating is not without some positive aspects; overraters exhibit high achievement, have high levels of self-esteem and social desirability, and tend to be less anxious (Brutus, Fleener, & McCauley, 1999; Goffin & Anderson, 2007; Nowack & Mashini, 2012). Nowack and Mashini (2012) suggested that these positive outcomes may be due to the fact that overraters tend to prioritize future achievement and selectively focus only on positive behaviors.

Somewhat surprisingly, recent research has found that positive outcomes may also be associated with underrating. While Eurich (2017) notes that underraters are blind to their strengths, underraters are perceived as more effective leaders and have more engaged employees (Zenger & Folkman, 2015). Although they rate themselves lower than observers, underraters tend to positively impact others, especially their subordinates, making them particularly effective leaders (Bratton, Dodd, & Brown, 2001; Zenger & Folkman, 2015). Underraters tend to be rated highest by their followers (e.g., Sosik & Magerian, 1999) and build the highest trust and organizational commitment (Sosik, 2001).

It is clear that SOA has important implications for leadership performance, and our goal is to evaluate these outcomes using an improved research methodology. All of the SOA research we reviewed used 360-degree feedback assessments, which vary greatly in accuracy and effectiveness (for a review, see Nowack & Mashini, 2012). While some suggest that others (e.g., superior, peer, direct report) can evaluate us more objectively than we evaluate ourselves (Eurich, 2017), most 360-degree feedback is only moderately objective. Raters, “may have limited ability or motivation to provide accurate and constructive feedback” (Bommer, Rubin, & Bartels, 2005, p. 103). Using MBA student performance does provide some advantages over past SOA research, which typically relies on subjective multi-source rating instruments. These tools vary in accuracy and effectiveness (e.g., Nowack & Mashini, 2012) due to raters’ lack of ability, training, and motivation (Bommer, Rubin, & Bartels, 2005). GPA and the in-basket are standardized, objective performance indicators in where all participants complete identical work. Performance is documented, and raters have ample time and opportunity to observe behaviors and performance. Furthermore, the assessment center does provide a realistic management performance scenario that predicts future managerial success (Gaugler, Rosenthal, Thornton, & Bentson, 1987; Howard, 1997; Whetzel, Rotenberry, & McDaniel, 2014).

In the current study, we utilized a managerial in-basket assessment center that provides numerous methodological improvements. First, the assessment center utilizes multiple, trained external raters who have no previous knowledge of the participants. Therefore, their ratings will be more objective and less biased, and interrater reliability is enhanced because any rating disagreements are discussed and settled with a focus on objective performance criteria.

Second, past SOA research tends to rely on subjective measures of leadership performance, rather than actual, objective performance. Performance outcomes are typically operationalized with a behavioral survey like the Multifactor Leadership Questionnaire (MLQ), which asks respondents to report on how frequently a leadership target displays certain leadership behaviors (e.g., Bass & Yammarino, 1991; Bratton et al., 2011). These are indirect performance

measures because the perception of the presence or absence of these behaviors is not an actual measure of leadership performance. In the assessment center, task performance is objectively evaluated in a standardized scenario where all participants complete the same tasks. Performance is documented with both written work and videotaped performance scenarios, so all raters have ample time and opportunity to observe leadership behaviors and performance. This standardization is not possible in workplace settings, and it significantly improves the validity of the outcomes.

If self-awareness is truly a key driver of leadership effectiveness, we would expect self-other agreement to be a better predictor of assessment center performance and GPA than either over- or underrating. Although self-awareness is considered essential for effective leadership (Eurich, 2018), underrating has been shown to have a significant positive impact on leadership outcomes, perhaps even moreso. Overall, these results do call into question whether it is rating accuracy or under-rating that is more important for leadership performance, and this is the question we seek to explore. Therefore, we expect that underraters will exhibit the highest levels of performance, followed by accurate raters, and then overraters.

3. Method

3.1. Participants

We collected data from students enrolled in a part-time MBA program at a regional university in the Pennsylvania State System of Higher Education. The sample was drawn from 176 students over a six-year period. Students were excluded if they did not complete the managerial in-basket prior to completing their degree. The final sample consisted of 129 participants: 47 women and 82 men. Ages ranged from 21 to 65 years, with an average age of 27.7 years.

3.2. Measures

3.2.1. Iliad Assessment Center. The Iliad Assessment Center (Bommer & Bartels, 1996) is a 145-minute simulation of a day in the life of a manager who has been away from work and returns to find the following tasks to complete: 1) Attend two 20-minute leaderless group discussion meetings, one about customer service initiatives, and the other to identify a potential CEO successor, 2) deliver a 3-minute persuasive speech about the participant's vision for the organization, and 3) work on an in-basket exercise in the role of a top manager in a publishing company. The Iliad has been validated and employed in other published studies (e.g., Rode, Arthaud-Day, Mooney, Near, Baldwin, Bommer, & Rubin, 2005).

The in-basket is a work document that includes the times and content of the leaderless group discussion meetings and speech times, administrative information (contact information, organizational charts, etc.), and a series of managerial tasks that require some action on the part of the participant (e.g., financial and operational decisions, communications, delegation). Participants must prioritize their workloads and make important decisions about their time and schedule since it is not possible to complete every activity in the time period. There are several components of the in-basket that contribute to the score that are also assessed, and these include work pace (number of memos completed), attention to detail (relative number of correct responses to items when heightened attention to detail was required), prioritization (how well the in-basket was prioritized by completing important memos), and writing quality (spelling and grammatical quality of written work as compared to others).

The speech and group discussions are recorded on video, and the in-basket materials are handwritten by participants. Following the simulation, all materials are sent to the Iliad Assessment Center for rating based on the presence and effectiveness of specific behaviors. The ratings were completed by current or former students in a master's of industrial psychology program. Two independent ratings were completed, and conflicts were resolved by collaboratively reviewing the recordings and coming to a consensus. Scores are provided for the following: Selection (effective evaluation of CEO candidates in a team setting), Customer Service (effective evaluation of different customer service initiatives in a team setting), In-Basket (materials participants must act upon during the assessment by sending emails, memos, etc.), Speech (verbal, non-verbal, and content-related aspects of a 3-minute persuasive speech), and Total Assessment Center Score (an equal-weighting combination of Selection, Customer Service, In-Basket, and Speech scores).

3.2.2. Rating Accuracy. The classification for over, under, or self-aware rating was made using a procedure well established in the literature by Atwater and Yammarino (1992). Rating accuracy was defined as the difference score between the self-ratings of participants and the actual scores assigned by trained raters. Before starting the Iliad Assessment, participants were asked to complete a self-assessment of their skills in comparison to their peers; this was operationalized as a percentage of the people they would outperform in Leadership Initiative, Decision Making, Teamwork, Communication, and Organizing. The actual score on each dimension was provided by the external Iliad

raters as part of the assessment center results. Participants were categorized into one of three groups by establishing a difference score between self and the average of other scores. Participants whose difference scores were half a standard deviation above the mean were considered over-raters. Participants whose difference score was half a standard deviation below the average were labeled underraters. Lastly, individuals with half a standard deviation above or below the mean were categorized as self-aware.

3.2.3. Graduate Grade Point Average (GGPA). GGPA was included as a measure of the academic performance of participants who complete the MBA program at our school. GPA was measured on a 4.0 scale; however, University policy indicates that students must maintain a GPA of 3.0 or higher to remain in good standing in the program. As a result, GPA included in this sample was restricted to a range of 3.0 - 4.0 with a mean of 3.65.

4. Results

To ensure that SOA membership was due to differences in self-perception rather than performance, we ran an ANOVA on self-ratings for each group. Results showed a significant group effect: $F=38.61$, $p<.01$, partial $\eta^2=.380$. Underraters had lower self-ratings than the Self-aware group, which in turn had lower self-rating than the Overrater group (all $p<.01$). Thus, we are confident that rating accuracy reflects how participants evaluate their selves.

Descriptive statistics are displayed in Table 1. For each performance measure, the Underrating group has a higher mean (and median) than the Self-aware group, which in turn has a higher mean (median) than the Overrating group. Skewness and kurtosis (now shown) are less than 2 in absolute value for all the performance indicators across groups, with the exception of the Self-aware group for GPA (4.35).

Table 1 Descriptive Statistics

Variable	Stat	All (N=129)	Rating Accuracy		
			Under-raters (N=33)	Self-aware (N=71)	Over-raters (N=25)
Graduate GPA	Mean (SD)	3.65 (.26)	3.74 (.21)	3.64 (.26)	3.52 (.29)
	Median	3.70	3.78	3.70	3.49
	Min/Max	3.0/4	3.25/4	3.0/4	3.10/3.95
ILIAD Total	Mean (SD)	52 (28)	76 (18)	54 (22)	16 (13)
	Median	52	76	52	16
	Min/Max	1/100	37/100	9/95	1/50
Self-Rating	Mean (SD)	67 (17)	52 (14)	68 (14)	83 (8)
	Median	69	50	70	84
	Min/Max	20/95	20/82	20/91	63/95
ILIAD Activity Selection	Mean (SD)	59 (28)	74 (21)	63 (25)	29 (22)
	Median	65	80	66	27
	Min/Max	1/100	17/99	6/100	1/79
Customer Service	Mean (SD)	56 (29)	61 (30)	59 (29)	38 (23)
	Median	54	76	66	35
	Min/Max	1/99	6/99	1/99	2/74
Speech	Mean (SD)	52 (26)	63 (25)	52 (23)	37 (27)
	Median	56	69	54	31
	Min/Max	2/92	7/92	5/92	2/89
In-Basket	Mean (SD)	48 (28)	69 (23)	48 (24)	17 (14)
	Median	47	74	47	12
	Min/Max	2/99	6/99	5/98	2/46
In-Basket Activity Work Pace	Mean (SD)	54 (28)	72 (22)	55 (25)	29 (23)
	Median	56	77	53	25
	Min/Max	2/100	13/100	6/99	2/82
Attention to Detail	Mean (SD)	56 (27)	66 (23)	57 (28)	39 (25)
	Median	57	69	62	42
	Min/Max	1/99	14/98	1/99	2/82
Prioritization	Mean (SD)	55 (28)	76 (26)	56 (23)	26 (21)
	Median	52	84	52	15
	Min/Max	4/100	10/100	10/100	4/63
Writing Quality	Mean (SD)	48 (28)	70 (25)	48 (24)	19 (15)
	Median	49	75	49	13
	Min/Max	2/100	9/100	4/97	2/57

Note. Mean and SD for the ILIAD are rounded to nearest integers.

Table 2 shows sample correlations between SOA and performance. Since SOA is an ordinal variable with three levels (-1 for under-raters, 0 for self-aware, and 1 for over-raters), we compute polyserial correlations, which are higher than Pearson correlations (.03 to .07). The polyserial correlations are all significant ($p < .01$) and negative. ILIAD Total is more strongly correlated with Rating Accuracy than is GPA. GPA is weakly correlated with the ILIAD Total and activity scores, which is not surprising since they measure different domains. The correlations among the four activity scores are significant, but low to moderate in magnitude, indicating discriminant validity for the ILIAD components.

Table 2 Correlations

Variable	1	2	3	4	5	6
1. Rating Accuracy						
2. Graduate GPA	-.31***					
3. ILIAD Total	-.77***	.32***				
4. Selection	-.57***	.23***	.65***			
5. Customer Service	-.27***	.10	.51***	.42***		
6. Speech	-.37***	.10	.49***	.26***	.11	
7. In-Basket	-.68***	.29***	.78***	.25***	.08	.21**

* $p < .1$, ** $p < .05$, and *** $p < .01$

Note. Column 1 lists polychoric correlations treating Rating Accuracy as an ordinal variable. Other entries are Pearson product-moment correlations. Rating Accuracy is coded: -1 (under-raters), 0 (self-aware), and 1 (over-raters).

We first ran a one-way MANOVA with Graduate GPA and ILIAD Total. Henze-Zirkler test rejects the null of multivariate normality for the Under-rater and Self-aware groups, but not for the Over-rater group ($p = .064$). Box's test does not reject the homogeneity of covariance matrices across the three groups ($M = 11.70$, $F = 1.89$, $p = .078$). Mardia (1971) indicated that deviation from multivariate normality has only a minor impact on Type I error (Mardia, 1971). The null hypothesis of no effects of Rating Accuracy is rejected: Wilks' $\Lambda = .485$, $F = 27.22$, $p < .01$, and partial $\eta^2 = .303$. As we establish joint significance, we follow up by univariate one-way ANOVAs (Table 3). Univariate tests for each performance criterion yield a significant F -statistic ($p < .01$). The explanatory power of SOA is low for GPA, while it explains slightly over half of the variation in ILIAD Total. Therefore, SOA is a powerful predictor of overall ILIAD performance.

Table 3 Univariate ANOVA Results

Variable	Parametric		Kruskal-Wallis X^2	Median Score X^2	Savage Score X^2
	F	R^2			
Graduate GPA	5.26	.08	9.30	3.77	1.50
ILIAD Total	65.97	.51	64.06	41.58	44.72
ILIAD Activity					
Selection	28.66	.31	35.41	25.00	19.32
Customer Service	5.95	.09	11.36	3.95	10.69
Speech	8.11	.11	14.51	8.58	13.74
In-Basket	40.07	.39	50.26	41.58	35.57
In-Basket Activity					
Work Pace	22.59	.26	32.51	28.77	23.22
Attention to Detail	7.78	.11	13.48	11.98	10.16
Prioritization	32.82	.34	41.66	31.08	36.25
Writing Quality	35.06	.36	45.99	36.10	37.68

Note. All F and X^2 statistics are significant ($p < .01$), except Median Score X^2 for Graduate GPA ($p = .15$), Customer Service ($p = .139$), and Speech ($p = .014$).

To determine which groups contribute to the overall difference, we implement the Tukey HSD procedure. For Graduate GPA, the only significant group difference is found between Under- vs. Over-raters (3.74 vs. 3.49). The Self-aware group does not achieve a significantly higher GPA than the Over-rater group. Cohen's d indicates a large effect for the difference between Under- and Over-raters in Graduate GPA. In contrast, the effects of rating accuracy are more pronounced on ILIAD Total by both statistical significance and effect-size measures. All the three pairwise comparisons are significant, suggesting that the three groups are ranked in order of performance: Under-raters > Self-aware > Over-raters. The values of d are much higher than those for Graduate GPA. On average, under-raters rank 22 percentiles higher than self-aware individuals.

Table 4 Pairwise Comparison by Tukey Method

Variable	Pair	Mean Difference	95% LB	95% UB	Cohen's <i>d</i>	SE of <i>d</i>
Graduate GPA	Under – Self	.10	-.04	.24	.41	.21
	Under – Over	.22**	.04	.40	.89	.28
	Self – Over	.12	-.04	.27	.44	.23
ILIAD Total	Under – Self	21.66**	1.70	32.63	1.03	.22
	Under – Over	59.70**	45.90	73.50	3.68	.43
	Self – Over	38.04**	25.93	5.14	1.88	.27
ILIAD Activity						
Selection	Under – Self	10.55	-3.73	24.84	.45	.21
	Under – Over	45.15**	27.17	63.12	2.09	.33
	Self – Over	34.59**	18.83	50.36	1.43	.25
Customer Service	Under – Self	1.92	-15.27	19.11	.07	.21
	Under – Over	22.90**	1.27	44.53	.85	.28
	Self – Over	20.98**	2.01	39.96	.76	.24
Speech	Under – Self	11.20	-3.70	26.09	.47	.21
	Under – Over	26.12**	7.38	44.86	1.00	.28
	Self – Over	14.92**	-1.52	31.36	.62	.24
In-Basket	Under – Self	21.13**	7.73	34.53	.90	.22
	Under – Over	52.19**	35.33	69.06	2.63	.36
	Self – Over	31.06**	16.27	45.85	1.44	.25
In-Basket Activity						
Work Pace	Under – Self	17.03**	2.28	31.77	.70	.22
	Under – Over	43.08**	24.53	61.63	1.90	.32
	Self – Over	26.05**	9.78	42.33	1.05	.24
Attention to Detail	Under – Self	8.65	-7.25	24.55	.33	.21
	Under – Over	26.94**	6.93	46.95	1.12	.28
	Self – Over	18.29**	0.74	35.84	.68	.24
Prioritization	Under – Self	20.16**	6.03	34.29	.85	.22
	Under – Over	49.82**	32.03	67.61	2.11	.33
	Self – Over	29.66**	14.06	45.26	1.33	.25
Writing Quality	Under – Self	21.71**	7.78	35.63	.89	.22
	Under – Over	50.77**	33.24	68.29	2.41	.35
	Self – Over	29.06**	13.69	44.43	1.31	.25

Note. **: $p < .025$ for Graduate GPA and ILIAD Total and $p < .0125$ for the ILIAD component and In-Basket component scores. The 95% lower and upper bounds are similarly adjusted for multiple comparisons to achieve the experiment-wise α of 5%.

Since the group effects are significant and large for ILIAD Total, we further investigate the source of group differences. Since ILIAD Total is an average of the component scores, an accumulation of small effects may result in a significant overall difference. Alternatively, a large effect may exist only in a few components. Therefore, it is imperative to identify the activities in which rating accuracy produces an effect. We run a one-way MANOVA with the four activity scores as dependent variables. Multivariate normality of data is rejected by the Henze-Zirkler test for the Self-aware ($p = .021$), but not for the Under- or Over-rater ($p = .211$ and $p = .074$, respectively). The equality of covariance matrices cannot be rejected by the Box's test ($M = 19.47$, $F = .92$, $p = .564$).

The multivariate null hypothesis of no effect of Rating Accuracy is again rejected: Wilks' $\Lambda = .419$, $F = 16.73$, $p < .01$. We obtain a slightly higher partial η^2 of .352. Table 4 shows that all the follow-up univariate ANOVAs are significant

($p < .01$). Post-hoc group comparisons (Tukey procedures) for the activity scores are presented in Table 4. Underraters and self-aware raters outperform overraters, and the standardized mean group differences have large effect sizes. The difference between the underraters and self-aware raters in Customer Service is marginal, statistically insignificant, and has a very small effect size ($p = .944$). Although the difference between underraters and self-aware raters is not significant in Selection and Speech, it is about 10 percentile points and has a meaningful effect size (medium effects). However, the In-Basket score exhibits a large, significant difference ($p < .01$) between the Underrater and Self-aware groups, and Cohen's d suggests a very large effect. The performance difference in In-Basket makes the greatest contribution to the significant differences between the Underrater and Self-aware groups in the ILIAD Total.

We further investigate if rating accuracy can predict significant and large differences in the In-Basket components: (1) Work Pace, (2) Attention to Detail, (3) Prioritization, and (4) Writing Quality. Neither multivariate normality nor homogeneity of covariance matrices is supported. MANOVA for the four In-Basket activities reveals a significant impact of Rating Accuracy: Wilks' $\Lambda = .595$, $F = 9.10$, $p < .01$, partial $\eta^2 = .228$. All F -statistics for univariate ANOVAs are significant (Table 3). As before, there are significant differences between Under- and Overraters and between the Self-aware and overraters in every component activity (Table 4). Underraters significantly outperform the self-aware raters on all components except for Attention to Detail. Kruskal-Wallis, Median score, and Savage score tests all indicate significant differences by rating accuracy (Table 3). The results for the In-Basket activities reinforce our finding that rating accuracy is an important predictor of performance. Specifically, underraters tend to outperform both the overraters and self-aware raters.

5. Discussion

In this study, we examined how SOA influenced leadership performance and academic (GPA) outcomes in a sample of MBA students. Overall, our results showed that SOA had a significant impact on all performance outcomes, explaining slightly over half of the total variance in assessment center performance. Underraters showed the highest levels of performance, regardless of how performance was measured (GPA and assessment center). As can be seen from Table 1, underraters exhibited the highest mean scores on every performance measure included in the study, followed by self-aware, and then overraters. Underraters significantly outperformed both self-aware and overraters on five of the ten outcomes measures, including ILIAD Total, In-Basket, Work-Pace, Prioritization, and Writing Quality. On the remaining five outcomes, GPA, Selection, Customer Service, Speech, and Attention to Detail, underraters significantly outperformed overraters. While the difference between underraters and self-aware raters was not significant for Selection and Speech, it was practically meaningful; approximately ten percentage points, which is considered a medium effect size. Thus, underraters significantly outperformed self-aware individuals on 50% of the outcomes, and overraters on 100% of the outcomes. Accurate raters did significantly outperform overraters on all performance measures except for GOA.

One thing is abundantly clear; overraters exhibited the lowest academic and assessment center performance. They were significantly outperformed by both self-aware and underraters on nearly every performance dimension. The largest performance differences were between under- and overraters. We conclude that overrating is not only bad for leadership effectiveness as previous research has shown (Zenger & Folkman, 2015), it is also bad for performance.

The most interesting findings from the study are the performance outcomes associated with underrating. Past research has suggested that self-aware leaders are more successful (e.g., Bass & Yammarino, 1991; Flaum, 2009), and that, "self-awareness is the single greatest predictor of leadership success" (Eurich, 2017, p. 154). Our results clearly suggest otherwise, as underraters outperformed self-aware individuals on every performance measure. Underrating has a far bigger impact on performance than self-awareness, and that these differences are both statistically and practically meaningful. As an example, underraters scored 22 percentiles higher than self-aware individuals on the assessment center. Our results raise questions about the use of accurate self-assessments as an indicator of leadership effectiveness.

The question does arise as to whether underrating represents a lack of self-awareness because it is technically a rating error. The degree of awareness and intention of the rater seems to be of importance to understand best the value of ratings, performance, and organizational outcomes. Underrating could stem from insecurity or a lack of self-confidence, and in this case, the individuals are unaware of their actual ability level. However, most of the research on underrating seems to indicate that it does not stem from a lack of self-awareness, but rather represents the actions of humble individuals (Sosik, 2001). Van Velsor, Taylor, and Leslie (1993) found that underraters were rated highest by direct reports on self-awareness. Zenger and Folkman (2015) noted that underraters were humble, had high personal standards, and had a continuous drive to improve. Humble individuals see themselves as a work in progress (Owens & Hekman, 2012) and feel as though they fall short of a standard they cannot reach (Ou, Tsui, Kinicki, Waldman, Xiao, & Song, 2014).

In this case, underraters do not lack self-awareness; they are keenly aware of their skills and abilities, but also believe they continually need to develop, evolve, and improve. This constant drive for improvement may result in higher levels of performance.

Similarly, overraters may be aware or unaware of their abilities. If they are aware of their lower abilities and intentionally inflate their skills, abilities, and performance, it would suggest they are deceptive (Atwater, Ostroff, Yammarino, & Fleenor, 1998). In this case, overrating is a form of self-promotion, which has been negatively related to transformational leadership and measures of satisfaction and effectiveness (Gardner & Cleavenger, 1998). Additionally, overraters may be narcissistic. Self-enhancement is a central aspect of narcissism (Grijalva & Zhang, 2016), and narcissists tend to overrate their attributes and abilities, such as leadership effectiveness, intelligence, physical attractiveness, openness, and honesty. The position of the individual may not always be one based on deception or negative intentions, as overraters may be entirely unaware of their low levels of skill, ability, and performance. In this case overrating would be unintentional. This blind spot is known as the Dunning-Kruger effect, a phenomenon whereby people are, “unskilled and unaware” (Kruger & Dunning, 1999, p. 1121).

6. Implications

These results are likely generalizable to organizations because assessment centers are predictive of future success. Assessment centers have proven to be effective tools for selection, promotion, training, and career development (Klimoski & Brickner, 1987), and are useful,” for predicting managerial success, across organizations and types of employees and for a number of purposes.” (p.245).

Therefore, organizations should be identifying these individuals in the selection and promotion processes, particularly for positions in leadership. Identifying and developing high potential leaders is a central concern of organizations, and these results call into question the belief that self-awareness is essential for leadership effectiveness. Employees have a preference for leaders who underrate themselves (e.g., Zenger & Folkman, 2015), and our results showed that underraters, not self-aware raters, demonstrate the highest levels of performance. This is a powerful combination of outcomes that should be leveraged by organizations.

7. Limitations and Future Directions

The current article had several limitations. First, the study included only MBA students in a simulated work environment. While in-basket assessments are predictive of future job performance, we did not have actual, objective work-related performance outcomes. Therefore, we do not know if these results would be replicated with objective measures of actual on-the-job performance. In addition, our sample is drawn from only one relatively small, public, regional MBA program. The majority of participants are Caucasian, from the region, and born in the United States. Additionally, research is needed on different samples from different institutions in order to generalize the results.

More work needs to explore why, and under what conditions, underestimating our abilities is beneficial. Grant (2021) noted that, “...we’re sometimes better off underestimating ourselves,” and so we need a better understanding of when this is the case. To this end, it is essential to identify the mechanisms that underlie underrating, not only to understand the root causes of underrating, but to use this knowledge to better understand SOA as a construct. Statistically speaking, it is a difference between self and other ratings, but what does this difference actually represent? To this end, it is crucial to determine the relationships between SOA and relevant individual characteristics, such as personality, self-esteem, self-monitoring, humility, deceptive tendencies, and narcissism. The attributes of underraters will be of great interest to anyone concerned with developing high-level performance skill sets individually or as an organization. Specifically, understanding why underraters tend to perform better has important implications for management education and management development.

References

- Ashford, S. J. (1989). Self-assessments in organizations: A literature review and integrative model, In Cummings, L.L. & Staw, B.W. (Eds.), *Research in organizational behavior*, JAI Press, Greenwich, CT, pp. 133–174.
- Atwater, L.E., Ostroff, C., Yammarino, F.J. & Fleenor, J.W. (1998). Self-other agreement: Does it really matter? *Personnel Psychology*, 51(3), 577-598.
- Atwater, L.E., & Yammarino, F.J. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership performance expectations? *Personnel Psychology*, 45, 275-279.
- Atwater, L.E. & Yammarino, F.J. (1997). Self-other rating agreement: A review and a model. *Research in Personnel and Human Resources Management*, 15, 121-174.

- Bass, B. & Yammarino, F. (1991). Congruence of self and others' leadership ratings of naval officers for understanding successful performance. *Applied Psychology: An International Review*, 40(4), 437-454.
- Bommer, W. H., & Bartels, L. K. (1996), *The Iliad Assessment Center*, Tichenor, Bloomington, IN.
- Bommer, W.H., Rubin, R.S., & Bartels, L.K. (2005). Assessing the unassessable: Interpersonal and managerial skills. In K. Martell & T. Calderon, T. (Eds.), *Assessment of Student Learning in Business Schools: Best Practices Each Step of the Way* (pp.103-120). Tampa, FL: AACSB.
- Bratton, B.K., Dodd, N.G., & Brown, F.W. (2011). The impact of emotional intelligence on accuracy of self-awareness and leadership performance. *Leadership and Organization Development Journal*, 32(2), 127-149.
- Brutus, S., Fleenor, J.W., & McCauley, C.D. (1999). Demographic and personality predictors of congruence in multi-source ratings. *Journal of Management Development*, 18(5), 417-435.
- Eurich, T (2017). *Insight: Why we're not as self-aware as we think, and how seeing ourselves clearly helps us succeed at work and in life*. New York, NY: Currency.
- Eurich, T. (2018). How to become more self-aware. *Harvard Business Review*. Retrieved from <https://hbr.org/ideacast/2018/06/how-to-become-more-self-aware>
- Flaum, J.P. (2009). When it comes to business leadership, nice guys finish first, available at: http://talentmgt.com/articles/view/new_study_finds_nice_guys_finish_first_when_it_comes_to_performance (accessed 12 December 2019)
- Fleenor, J., Smither, J, Atwater, L., Braddy, P., & Sturm, R. (2010). Self-other rating agreement in leadership: A review. *The Leadership Quarterly*, 21(6), 1005-1034.
- Gardner, W.L. & Cleavenger, D. (1998). The impression management strategies associated with transformational leaders at the world-class level. *Management Communication Quarterly*, 12(1), 3-41.
- Gaugler, B.B., Rosenthal, D.B., Thornton, G.C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, 72(3), 493-511.
- George, B, Sims, P., McLean, N., & Mayer, D. (2007). Discovering your authentic leadership. *Harvard Business Review*, 85(22), 129 – 138.
- Grant, A. (2021). *Think Again: The Power of Knowing What You Don't Know*. New York, NY: Viking.
- Goffin, R., & Anderson, D. (2007). The self-rater's personality and self-other disagreement in multisource performance ratings: Is disagreement healthy? *Journal of Managerial Psychology*, 22(3), 271-289.
- Grijalva, E. & Zhang, L. (2016). Narcissism and self-insight: A review and meta-analysis of narcissists' self-enhancement tendencies. *Personality and Social Psychology Bulletin*, 42(1), 3-24.
- Harris, M.M. & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, 41, 43-62.
- Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality*, 12(5), 13-52.
- Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology*, 40, 243 – 260.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121-1134.
- Mabe, P.A. & West, S.G. (1982). Validity of self-evaluation ability: A review and meta-analysis. *Journal of Applied Psychology*, 67(3), 280-296.
- Mardia, K.V. (1971). The effect of non-normality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, 58(1), 105-121.
- Moshavi, D., Brown, F., & Dodd, N. (2003). Leader self-awareness and its relationship to subordinate attitude and performance. *Leadership and Organization Development Journal*, 24(7), 407-481.
- Nowack, K.M., & Mashihhi, S. (2012). Evidence-based answers to 15 questions about leveraging 360-degree feedback. *Consulting Psychology Journal: Practice and Research*, 64(3), 157-182.
- Ou, A.Y., Tsui, A.S., Kinicki, A.J., Waldman, D.A., Xiao, Z., & Song, L.J. (2014). Humble chief executive officers' connections to top management team integration and middle managers responses. *Administrative Science Quarterly*, 59(1), 34-72.
- Owens, B.P. & Hekman, D.R. (2012). Modeling how to grow: An inductive examination of humble leader behaviors, contingencies, and outcomes. *Academy of Management Journal*, 55(4), 787-818.
- Rode, J.C., Arthaud-Day, M.L., Mooney, C.H., Near, J.P., Baldwin, T.T., Bommer, W.H., & Rubin, R.S. (2005). Life satisfaction and student performance. *Academy of Management Learning and Education*, 4(4), 421-433.
- Sosik, J.J. (2001). Self-other agreement on charismatic leadership: Relationships with work attitudes and managerial performance. *Group and Organization Management*, 26(4), 484-511.

- Sosik, J. J., & Megerian, L. E. (1999). Understanding leader emotional intelligence and performance. *Group Organization Management*, 24(3), 367-390.
- Vecchio, R. & Anderson, R. (2009). Agreement in self–other ratings of leader effectiveness: The role of demographics and personality. *International Journal of Selection and Assessment*, 17(2), 165-179.
- Whetzel, D.L., Rotenberry, P.F., & McDaniel, M.A. (2014). In- basket Validity: A systematic review. *International Journal of Selection and Assessment*, 22(1), 62-79.
- Woo, S.E., Sims, C.S., Rupp, D.E., & Gibbons, A.M. (2008). Developmental engagement with and following developmental assessment centers: Considering feedback favorability and self-assessor agreement. *Personnel Psychology*, 61(4), 727-759.
- Yammarino, F., & Atwater, L. (1993). Understanding self-perception accuracy: Implications for human resource management. *Human Resource Management*, 32(2/3), 231-247.
- Yammarino, F., & Atwater, L. (1997). Do managers see themselves as others see them? Implications for self-other rating agreement of human resources management. *Organizational Dynamics*, 25(4), 35 – 44.
- Zenger, J. (2014). The singular secret for a leader's Success: Self-awareness. *Forbes*. Retrieved from <http://www.forbes.com/sites/jackzenger/2014/04/17/the-singular-secret-for-a-leaders-success-self-awareness/#4571114177e4> (Accessed 5 November 2019).
- Zenger, J., & Folkman, J. (2015). We like leaders who underrate themselves. *Harvard Business Review*. Retrieved from <https://hbr.org/2015/11/we-like-leaders-who-underrate-themselves>