Technical Report 1386

Effects of Group Size on Predictive Validity of Peer Assessments of Leadership

Elizabeth R. Uhl
Melissa J. Glorioso
U.S. Army Research Institute

September 2020

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

Department of the Army
Deputy Chief of Staff, G1

Authorized and approved:

MICHELLE L. ZBYLUT,
Ph.D. Director

Technical review by

Jayne Allen, U.S. Army Research Institute

NOTICES

DISTRIBUTION: This Technical Report has been submitted to the Defense Information Technical Center (DTIC). Address correspondence concerning ARI reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPE-ARI-ZXM, 6000 6th Street Building 1464 / Mail Stop: 5610), Fort Belvoir, VA 22060-5610.

FINAL DISPOSITION: Destroy this Technical Report when it is no longer needed. Do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Technical Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 04-09-2020	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* Nov 2016 – Oct 2017
4. TITLE AND SUBTITLE Effects of Group Size on Predictive Validity of Peer Assessments of Leadership		**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER** 622785
6. AUTHOR(S) Elizabeth R. Uhl & Melissa J. Glorioso		**5d. PROJECT NUMBER** A790 **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U. S. Army Research Institute for the Behavioral & Social Sciences 6000 6TH Street (Bldg. 1464 / Mail Stop 5610) Fort Belvoir, VA 22060-5610		**8. PERFORMING ORGANIZATION REPORT NUMBER** Technical Report 1386
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U. S. Army Research Institute for the Behavioral & Social Sciences 6000 6TH Street (Bldg. 1464 / Mail Stop 5610) Fort Belvoir, VA 22060-5610		**10. SPONSOR/MONITOR'S ACRONYM(S)** ARI **11. SPONSOR/MONITOR'S REPORT NUMBER(S)** Technical Report 1386

12. DISTRIBUTION/AVAILABILITY STATEMENT: Distribution Statement A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Subject Matter POC: Elizabeth Uhl

14. ABSTRACT

Peer assessments have demonstrated predictive validity for many military outcomes, including academic standing and leadership, success in training, and combat effectiveness. Though military training is often conducted at the squad or platoon level, little research has examined the effects of group size on the reliability and predictive validity of peer assessments for performance outcomes. The present study used archival data to examine the impact of group size on the interrater reliability and the predictive validity of peer rankings for leadership performance. Average peer rankings at the squad and platoon level were examined as predictors of performance outcomes for 191 junior Army leaders. Results indicated that platoon-level peer rankings tended to be better predictors of instructor-rated leadership scores in a garrison environment, though findings were mixed when examining leadership in field training exercises.

15. SUBJECT TERMS
Peer assessment, peer ranking, group size, leadership assessment, leader performance

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Dr. Jennifer S. Tucker
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	Unlimited Unclassified	20	**19b. TELEPHONE NUMBER** 706-545-2490

Technical Report 1386

Effects of Group Size on Predictive Validity of Peer Assessments of Leadership

Elizabeth R. Uhl
Melissa J. Glorioso
U.S. Army Research Institute

Fort Benning Research Unit
Jennifer S. Tucker, Chief

September 2020

EFFECTS OF GROUP SIZE ON PREDICTIVE VALIDITY OF PEER ASSESSMENTS OF LEADERSHIP

EXECUTIVE SUMMARY

Research Requirement:

Peer assessments are used in a variety of contexts throughout the U.S. Army. The current study examined the impact of group size (i.e., squad vs. platoon) on the predictive validity of peer assessments.

Procedure:

Archival data from junior Army leaders were examined to determine the effects of group size on the reliability and predictive validity of peer rankings for leadership performance. This was done by analyzing the relationships between squad-level and platoon-level peer rankings and instructor-graded leadership scores in an intensive military training setting.

Findings:

The analyses suggest that platoon-level peer rankings tended to be stronger predictors of leadership scores than squad-level peer rankings. Platoon-level peer rankings contributed unique variance above and beyond squad-level rankings for two out of the three leadership performance measures examined.

Utilization and Dissemination of Findings:

These findings can be used to guide decisions about peer assessment processes, such as, at what level (e.g., squad vs. platoon) peer assessments are conducted. However, it is important to consider the practical implications, such as the time to conduct the assessments and compile the data, of conducting peer assessments at platoon vs. squad levels, in addition to considering the statistical implications.

EFFECTS OF GROUP SIZE ON PREDICTIVE VALIDITY OF PEER ASSESSMENTS OF LEADERSHIP

CONTENTS

LIST OF TABLES

Effects of Group Size on Predictive Validity of Peer Assessments of Leadership

Introduction

In the U. S. Army, peer assessments are currently used in a variety of formats and contexts. For example, in some basic training units, trainees are asked to rate their squad members on the Army values (U.S. Department of the Army, 2012), which include characteristics such as loyalty, duty, and respect. These new Soldiers are also asked to nominate the top and bottom five performers in their platoon. The results from these peer assessments are used in individual Soldier counseling for developmental purposes. Further, a senior leader course requires senior non-commissioned officers to complete a 360-degree assessment. This includes a peer assessment, as well as self, subordinate, and supervisor assessments, and provides specific feedback to individuals about areas for self-development. Another way the U.S. Army has used peer assessments is during junior officer training, in which peer assessments serve as a developmental tool, a graded assignment, and a supplement to instructor observations. Clearly then, peer assessments are perceived as valuable across different populations and across different use cases.

Studies have demonstrated that peer assessments are predictive of a range of critical military outcomes including combat performance (Levi, Torrance, & Pletts, 1954; Williams & Leavitt, 1947;), leadership (Reynolds, 1966; Wherry & Fryer, 1949), officer performance (Hollander, 1965), career advancement (Amir, Kovarsky, & Sharan 1970), ranger performance (Downey, 1976; Gilbert & Downey, 1978), and special forces training outcomes (Zazanis, Zaccaro, & Kilcullen, 2001). In a meta-analysis of the correlation between peer assessments and performance, Norton (1992) analyzed data from 32 studies (72% had a military sample) to examine the predictive validity of these ratings and found a mean correlation of $r = 0.69$ between peer assessment and performance.

One reason that peer assessments may be a good predictor of performance is that peers are often privy to behavior that an instructor does not see. Students may behave differently when in the presence of their instructor because they realize they are being evaluated. However, as students spend more time with each other than with their instructor, peers may experience a greater variety of interactions with each other which are less influenced by the need to be on their best behavior (Norton, 1992). Research suggests that peer assessments may be superior to those of supervisor assessments because the greater variety of interactions allows for finer distinctions to be made when evaluating various criteria (Schwarzwald, Koslowsky, & Mager-Bibi, 1999). This may be especially true in a military training environment where Soldiers are required to live in barracks and therefore spend time together during both training and rest periods, which may allow for more accurate assessments after short periods of time. Under these living conditions, peers would observe a comparably more extensive sample of an individual's behavior than is possible for supervisors to observe.

One's familiarity and knowledge of one's peers may vary within a peer group (e.g., squad or platoon). Therefore, group size may be an important consideration for predicting performance, particularly in the military context where training is often organized around the squad and

platoon. Army training is often conducted by companies consisting of 60-200 Soldiers organized into 3-4 platoons, with training or training exercises often occurring at the platoon or squad level. A squad consists of 4-10 Soldiers and a platoon consists of several squads (approximately 30-40 Soldiers). Because of the difference in the number of personnel, the level (i.e. squad or platoon) at which peer assessments are conducted could affect the overall reliability and predictive validity of the peer assessments.

In an early review of peer assessment research, Downey and Duffy (1978) noted that small group ratings may produce less reliable and valid scores. Despite this concern, little research has examined the effects of group size on the predictive validity or reliability of peer assessments. Most of the extant research on the effects of group size concentrates on the use of peer nomination or rating methods, leaving the effects of group size on peer rankings largely understudied. For example, Downey (1976) compared squad ratings with platoon nominations and found they were highly correlated but that platoon nominations were both more reliable and more predictive of job performance. However, measurement method (i.e., ratings and nominations) was confounded with group size, which limited the interpretability of the results. Research to examine the effects of group size is needed in order to delineate the most efficient and effective level of peer assessment in the military.

Impact of Number of Raters on Interrater Reliability and Predictive Validity

Peer evaluations are often used for summative assessment purposes. One concern that comes with multiple raters, especially in the context of assigning grades or scores, is interrater reliability. Higher interrater reliability indicates lower measurement error (Cohen, Cohen, West, & Aiken, 2003). In a meta-analysis of multisource performance ratings, Conway and Huffcutt (1997) examined the reliability of peer assessments across 26 samples and found an average interrater reliability of ICC = .37 for peer assessments. However, scant research has examined the impact of the number of raters on interrater reliability (Sung, Chang, Chang, & Yu, 2010). Sung et al. (2010) found that when rating individual performance, reliability tends to increase with the number of raters and that an acceptable level of reliability (≥ 0.70) was reached with 3-4 raters, and increased with more raters. Further research has suggested that 10 raters can be sufficient for establishing interrater agreement (LeBreton & Senter, 2008). Multiple peer ratings should be statistically more reliable than a single peer rating (Norton, 1992; Wherry & Fryer, 1949). Based on these findings, squad-level ratings are expected to have acceptable levels of reliability, though platoon-level peer assessments are expected to have higher interrater reliability than squad-level peer assessments, as they include more raters.

Hypothesis 1: Platoon-level peer rankings will have higher interrater reliability than squad-level peer rankings.

In one of the few studies to specifically examine the impact of the number of raters on the predictive validity of peer assessments of individual performance, Sung et al. (2010) found that validity reached an acceptable level with 3 or 4 raters but that validity continued to increase as the number of raters increased. Based on this finding, we expect platoon-level peer rankings to be better predictors of leadership performance than squad-level peer rankings.

Hypothesis 2: Platoon-level peer rankings will be a stronger predictor of leadership performance than squad-level peer rankings.

In order to address these gaps in the literature, the current study examined archival data from junior Army leaders to determine the effects of group size on the reliability and predictive validity of peer rankings for leadership performance. This was done by analyzing the relationships between squad-level and platoon-level peer rankings and instructor graded leadership scores in an intensive military training setting.

Method

Participants

Archival data of peer rankings and leadership performance were examined for 191 Soldiers from a 12-week intensive junior officer training course. Data from one class, including four platoons, each with four squads, for a total of 16 squads, was examined. Information on gender, age, and prior military experience was not collected.

Measures

Squad-level peer rankings

After six weeks in the course, Soldiers ranked each member of their squad, excluding themselves, from one (top performer) to *i* (bottom performer). Lower numbers (i.e., closer to one) indicate better rankings. Soldiers were also asked to answer two questions aimed at examining trust about each person they ranked: "Would you go to combat with this person?" and "Would you share a foxhole with this person?" An individual's rank score was determined by taking the average of all peer rankings (Squad Time 1, M = 7.52 raters, SD = 1.61, range 2-16 raters). After eight weeks in the course, Soldiers completed a second squad-level peer ranking using the same procedure as the Time 1 ranking (Squad Time 2, M = 4.73 raters, SD = 3.10, range 1-10 raters).

Platoon-level peer rankings

After 10 weeks in the course, Soldiers completed a platoon-level peer ranking. Soldiers ranked each member of their platoon, excluding themselves, from one to *i* (Platoon, M = 17.94 raters, SD = 9.87, range 3-33 raters). Soldiers were also asked to answer two questions about each person they rank: "Would you go to combat with this person?" and "Would you share a foxhole with this person?"

Garrison leadership scores

The majority of the course is spent in the garrison environment (i.e. the 'home base' environment), in which Soldiers take classes and conduct daily training. Each Soldier served in a garrison leadership position for one week on one or two occasions during the course, and leadership performance was graded by instructors for each week (i.e., Garrison Leadership 1 and

Garrison Leadership 2). Garrison Leadership scores were out of 330 points. As Garrison Leadership 2 scores were only available for 40 Soldiers, they were not included in the current analyses.

Field leadership scores

There are several field exercises in which Soldiers execute squad or platoon-level missions in an environment intended to more closely represent battlefield conditions. Each Soldier performed as a squad leader during a field exercise on two occasions. Multiple iterations of squad or platoon exercises were completed over several days, with individual leadership opportunities lasting 30-60 minutes. Leadership performance was graded for each field exercise by instructors (i.e., Field Leadership 1 and Field Leadership 2). Field Leadership scores were out of 330 points.

Data Analyses

Interrater reliability

Krippendorff's alpha (α) was calculated using an SPSS macro (Hayes & Krippendorff, 2007) to examine the interrater reliability of squad and platoon peer rankings. Krippendorff's alpha was used because, unlike other measures of interrater reliability, it can be used with a varying number of raters and has the ability to handle missing data.

Predictive Validity

Hierarchical regression analyses were conducted to examine the effects of group size (i.e., squad vs. platoon) on the predictive validity of peer rankings on performance. Separate regression analyses were conducted for each of the outcome variables: Garrison Leadership 1 scores, Field Leadership 1 scores, and Field Leadership 2 scores.

Results

Correlations among peer rankings and leadership scores (see Table 1), suggest that both squad- and platoon-level peer rankings were moderately correlated with instructor-graded leadership scores for most field and garrison relationships.

Table 1

Means, standard deviations, and correlations

Variable	1	2	3	4	5	6	M	SD	N
1. Squad T1	-						4.86	1.84	190
2. Squad T2	.68**	-					4.60	2.01	113
3. Platoon	.72**	.71**	-				18.80	7.27	191
4. Garrison	-.19**	-.19*	-.35**	-			280.12	12.34	191
5. Field 1	-.28**	-.20*	-.27**	.18*	-		268.78	19.55	182
6. Field 2	-.23**	-.10	-.27**	.16*	.35**	-	278.40	21.86	190

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

A paired sample *t*-test showed that the two squad-level peer rankings were not significantly different from each other, $t\ (112) = 0.66$, $p = 0.51$. To avoid redundancies in the model, only Squad T1 peer rankings were included in further analyses, as there were fewer missing cases for Squad T1 ($n = 190$) than for Squad T2 ($n = 113$) peer assessments.

Interrater Reliability

Krippendorff's alpha (Krippendorff, 1970) was calculated to measure the interrater reliability of the peer rankings. Interrater reliabilities were $\alpha = 0.41$ for squad-level rankings and $\alpha = 0.39$ for platoon-level rankings. These reliability findings were consistent with the results of Conway and Huffcutt's (1997) meta-analysis. These results did not support Hypothesis 1 that platoon-level rankings would have higher interrater reliability than squad-level rankings. The interrater reliability of squad-level rankings was slightly higher than that of platoon-level rankings, but it is unclear from the available data if this represents a significant difference.

Predictive Validity

A hierarchical regression analysis was conducted to examine the predictive validity of Squad T1 and Platoon rankings on Garrison Leadership 1 scores (see Table 2). Tests of multicollinearity were within acceptable levels, Tolerance = .48, VIF = 2.07. The results of the hierarchical regression analysis indicated that the overall model was significant when both of the variables were included and that the platoon-level variable was a better predictor of Garrison Leadership 1 than the squad-level variable. That is, when platoon-level peer rankings were added to the model, squad-level peer rankings did not significantly contribute to the prediction of garrison leadership. The negative relationship between Platoon rankings and Garrison Leadership 1 indicates that as peer rankings decrease in numerical value (i.e. get closer to 1, which is the top rank), leadership scores increase.

Table 2

Hierarchical Regression Analysis of Peer Rankings as Predictors of Garrison Leadership 1 Scores

Variable	*B*	95% CI for *B*		*SE B*	*β*	R^2	ΔR^2
		LL	*UL*				
Step 1						.04	.04**
Constant	286.42	281.49	291.36	2.50			
Squad T1	-1.29**	-2.24	-0.34	0.48	-0.19		
Step 2						.13	.09**
Constant	290.11	285.12	295.09	2.53			
Squad T1	0.82	-0.48	2.12	0.66	0.12		
Platoon	-0.74**	-1.07	-0.41	0.17	-0.44		

Note. CI = Confidence interval; *LL* = lower limit; *UL* = upper limit.
* $p < .05$. ** $p < .01$.

A hierarchical regression analysis was conducted to examine the predictive validity of Squad T1 and Platoon rankings on Field Leadership 1 scores (see Table 3). Tests of multicollinearity were within acceptable levels, Tolerance = .49, VIF = 2.04. While the overall model was significant, when Squad T1 and platoon peer rankings were included in Step 2 of the model, neither significantly contributed to the prediction of Field Leadership 1 scores. When examined independently, both Squad T1 (B = -2.98, SE = 0.77, $p < .01$) and platoon peer rankings (B = -0.75, SE = 0.20, $p < .01$) predicted Field Leadership 1 scores.

Table 3

Hierarchical Regression Analysis of Peer Rankings as Predictors of Field Leadership 1 Scores

Variable	*B*	95% CI for *B*		*SE B*	*β*	R^2	ΔR^2
		LL	*UL*				
Step 1						.08	.08**
Constant	283.09	275.23	290.94	3.98			
Squad T1	-2.98**	-4.50	-1.45	0.77	-0.27		
Step 2						.09	.01
Constant	285.29	277.00	293.59	3.98			
Squad T1	-1.74	-3.91	0.43	1.10	-0.16		
Platoon	-0.44	-1.00	0.11	0.28	-0.16		

Note. CI = Confidence interval; *LL* = lower limit; *UL* = upper limit.
* $p < .05$. ** $p < .01$.

A hierarchical regression analysis was conducted to examine the predictive validity of Squad T1 and Platoon rankings on Field Leadership 2 scores (see Table 4). Tests of multicollinearity were within acceptable levels, Tolerance = .48, VIF = 2.09. Results of the regression analyses indicated that the overall model was significant, and that the platoon-level variable was a better predictor of Field Leadership 2 scores than the Squad T1 variable. This was consistent with the results presented in Table 2: platoon-level peer rankings were better predictors of leadership than squad-level peer rankings.

Table 4

Hierarchical Regression Analysis of Peer Rankings as Predictors of Field Leadership 2 Scores

Variable	*B*	95% CI for *B*		*SE B*	*β*	R^2	ΔR^2
		LL	*UL*				
Step 1						.05	.05**
Constant	291.92	283.21	300.63	4.42			
Squad T1	-2.75**	-4.43	-1.08	0.85	-0.23		
Step 2						.08	.02*
Constant	294.95	285.84	304.06	4.62			
Squad T1	-0.94	-3.34	1.46	1.22	-0.08		
Platoon	-0.63*	-1.24	-0.03	0.31	-0.21		

Note. CI = Confidence interval; *LL* = lower limit; *UL* = upper limit.
* $p < .05$. ** $p < .01$.

These results provide partial support for Hypothesis 2. For the Garrison Leadership 1 and Field Leadership 2 measures, platoon-level rankings were better predictors than squad-level rankings. However, for Field Leadership 1, both measures were equally predictive.

Exploratory Analyses

The number of raters for each peer assessment varied, from 1 – 16, with an average of 7.66 (*SD* = 2.27) for Squad T1 peer assessments at Week 6 and from 3 – 33 with an average of 17.94 (*SD* = 9.87) for platoon level rankings at Week 10. Therefore, it is difficult to discern whether the difference in predictive validity is due to group size, group closeness, or timing. To begin to address this question, we examined the input of squad members compared to all platoon members at Week 10. By holding time constant, we were able to further examine the impact of group closeness without the noise from timing acting as a confounding variable. The number of squad raters at Week 10 varied, from 1 – 9 with an average of 4.09 (*SD* = 2.3). Not surprisingly, the squad-level and platoon-level peer rankings at Week 10 were strongly correlated, $r = .83$, $p < .001$. However, multicollinearity indicators were within acceptable levels, Tolerance = 0.31 – 0.34; VIF = 2.98 – 3.22). When both platoon and squad level variables were included in the model, platoon-level peer assessments significantly contributed to the regression model while squad-level peer assessments did not for both Field Leadership 1 and Field Leadership 2. The change in R^2 was significant when the platoon-level peer assessment was included for both Field

Leadership 1 and Field Leadership 2. For garrison leadership, though the overall model was significant, when both variables were included in the model neither significantly contributed to the model, though the impact of platoon-level peer assessments was trending towards a significant relationship. For full regression results, see Table 5.

These findings suggest that platoon level rankings may have greater predictive validity than squad rankings when timing is not a confound. While platoon level rankings do increase predictive validity, the increase is small (ΔR^2 range from .02 - .04). This is not unexpected because of the high correlation between these variables. Thus, squad level rankings may be acceptable when examining peer leadership.

Table 5

Hierarchical regression analyses of peer rankings at Time 3 by squad and platoon

Variable	*B*	95% CI for *B*		*SE B*	*β*	R^2	ΔR^2
		LL	UL				
			Garrison Leadership				
Step 1						.11	.11**
Constant	286.91	283.56	290.26	1.70			
Squad T3	-.43**	-0.61	-0.25	.09	-.33		
Step 2						.12	.02
Constant	290.08	285.19	294.98	.16			
Squad T3	-.19	-0.51	0.13	.16	-.15		
Platoon	-.37	-0.79	0.05	.21	-.22		
			Field Leadership 1				
Step 1						.04	.04**
Constant	275.02	269.87	280.47	2.77			
Squad T3	-.41**	-0.71	-0,10	.15	-.20		
Step 2						.07	.03*
Constant	281.97	274.08	289.85	4.00			
Squad T3	.10	-0.41	0.62	.26	.05		
Platoon	-.81*	-1.48	-0.14	.34	-.30		
			Field Leadership 2				
Step 1						.03	.03*
Constant	285.16	278.95	291.37	*3.15*			
Squad T3	-.41*	-0.74	-0.07	.17	-.17		
Step 2						.07	.04**
Constant	294.11	285.13	303.09	4.55			
Squad T3	.26	-0.33	0.86	.30	.11		
Platoon	-1.05**	-1.82	-0.28	.39	-.35		

Note. CI = Confidence interval; *LL* = lower limit; *UL* = upper limit.
* $p < .05$. ** $p < .01$.

Discussion

Archival data were analyzed to assess the effects of group size on the interrater reliability and predictive validity of peer rankings for leadership performance in a military training context. The interrater reliabilities for Squad T1 and platoon rankings were similar to the average

interrater reliability for peer evaluations in Conway and Huffcutt's (1997) meta-analysis. The slight difference (0.02) in reliability between Squad T1 and platoon assessments is not enough to draw a conclusion about the effect of group size on interrater reliability and it was in the opposite direction predicted by Hypothesis 1. Interestingly, the interrater reliability of Squad T2 (Krippendorf's $\alpha = 0.31$) was lower than Squad T1 (Krippendorf's $\alpha = 0.41$) and there were typically fewer raters for Squad T2 ($M = 4.73$) than for Squad T1 ($M = 7.52$).

Squad- and platoon-level peer rankings were highly correlated with one another ($r = 0.72$). The results of hierarchical regression analyses demonstrated that platoon-level peer rankings tended to be stronger predictors of leadership scores as they contributed unique variance above and beyond squad-level rankings for two out of three leadership performance measures, providing partial support for Hypothesis 2.

These findings suggest that platoon-level peer assessments can be better predictors of leadership scores than squad-level peer assessments. There are a number of reasons why this may be the case. In statistical analyses, larger sample sizes generally produce less error and more precise results (Cohen, 1988). The same principles are likely at work here. For example, as the number of raters increases, the impact of outlier scores decreases. Further, platoon-level peer rankings may mitigate the effects of bias because of the increased number of rankings available. A consistent concern with the use of peer assessments is that they are a popularity contest and may reflect friendship biases rather than actual performance. Though both squad- and platoon-level rankings may be influenced by personal bias, with a larger sample size, platoon rankings would be expected to be less biased. In a platoon-level assessment, there may be more raters who do not have personal experience with the individual being rated; therefore, they may be more likely to focus their rankings only on behaviors relevant to the assessment. With that said, a central advantage of peer evaluations is that they allow instructors insight into students' performance and behavior outside of an instructional context. The quality of those insights are likely to depend on the level of personal contact and familiarity a rater has with a given ratee. It is therefore reasonable to assume that squad members would be in the best position to provide quality feedback as long as the previously mentioned biases were mitigated. This is an important consideration for future research on this issue. Research should consider ways to potentially reduce bias in peer evaluations, such as using peer evaluations for only developmental purposes. Further, finding the right balance between quality and quantity of peer evaluation data will be an important aspect of future peer evaluation research.

In general, the results suggest that platoon-level peer rankings are more predictive of instructor-rated leadership scores than squad-level rankings. When applying these findings, it is important to consider the purpose of the peer assessment. Though platoon-level peer assessments are more predictive than squad-level assessments, the increase is small (ΔR^2 range from .02 - .04). While this might be an important advantage when conducting research, this advantage may not translate to practical application. That is, there might be additional resource costs to conducting platoon-level peer assessments that are unequal to the added benefit of collecting this data at the platoon level rather than the squad level.

Though these findings suggest that platoon-level peer rankings are more predictive of instructor-rated leadership scores than squad-level rankings, there are limitations to the current

study. In the current study, peer rankings were examined as predictors of instructor graded leadership scores. Instructor leadership scores were employed as the criterion for this study because of the nature of the sample. However, using peer rankings to predict instructor scores may not account for some of the expected advantages of peer assessments. For example, one of the justifications for using peer assessments is that peers are expected to see behaviors that evaluators do not see (Hollander, 1954). In addition, the behaviors seen by evaluators are subject to demand characteristics – when someone knows they are being evaluated, they may perform differently than when the evaluator is not present (Norton, 1992). Further, the peer assessment is a global assessment of performance in the course, it is not specific to any individual characteristic. Peer assessments of general performance may not predict leadership scores as well as peer assessments of leadership might. Understanding the purpose of the peer assessment is vital to determining the most appropriate criterion (O'Donnell & Topping, 1998), and thus should be taken into consideration.

There are also limitations due to the use of archival data rather than conducting experimental research. First, the platoon-level peer ranking was the last assessment conducted in the course. Therefore, it is possible that the greater predictive validity of the platoon rankings is due to the additional time the Soldiers spent together. Norton (1992) found that the validity of peer assessments was highest when peers had adequate time to become familiar with each other's skills, abilities, and qualifications, though Hollander (1957) found that peer ratings after one week were highly correlated with later peer ratings. However, comparison of platoon-level to squad-level means at Week 10 suggest that platoon level rankings have an advantage over squad-level rankings when compared at the same time point. A second limitation of using archival data is that the number of raters in each group was not precisely controlled, so we cannot draw conclusions about optimal group size.

These limitations suggest several avenues for future research. Ideally, the criteria selected for future research would conceptually match the focus of the peer assessments. It would also be informative to have multiple measures of the criteria, from different sources, not just an instructor or evaluator. Experimental research could control for the number of individuals in each group in order to assess whether an "optimum" group size exists, or if there is a point of diminishing returns for increasing predictive validity. Future research could also counterbalance the timing of small group and large group assessments. This would help to reduce the influence of potentially confounding variables identified in the current study.

References

Amir, Y., & Kovarsky, Y. (1970). Peer nominations as a predictor of multistage promotions in a ramified organization. *Journal of Applied Psychology, 54(5),* 462-469. https://doi.org/10.1037/h0029919

Cohen, J. (1988). *Statistical power analysis for the behavioral science* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*(4), 331-360. https://doi.org/10.1207/s15327043hup1004_2

Downey, R. G. (1976). *Associate nominations in the U. S. Army Officer Training Environment: The Ranger Course*. (ARI Research Problem Review 79-8). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. ADA076646).

Downey, R. G., & Duffy, P. J. (1978). *Review of Peer Evaluation Research.* (ARI Technical Paper 342). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. ADA 061780).

Gilbert, A. C. F., & Downey, R. G. (1978). *Validity of peer ratings obtained during Ranger training.* (ARI Technical Paper 344). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. ADA061576).

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communications Methods and Measures, 1*(1), 77-89. https://doi.org/10.1080/19312450709336664

Heyman, J. E., & Sailors, J. J. (2011). Peer assessment of class participation: applying peer nomination to overcome rating inflation. *Assessment & Evaluation in Higher Education, 36*(5), 605-618. https://doi.org/10.1080/02602931003632365

Hollander, E. P. (1954). Buddy ratings: military research and industrial applications. *Personnel Psychology, 7,* 385-393. https://doi.org/10.1111/j.1744-6570.1954.tb01607.x

Hollander, E. P. (1957). The reliability of peer nominations under various conditions of administration. *Journal of Applied Psychology, 41(2),* 85-90. https://doi.org/10.1037/h0047765

Hollander, E. P. (1965). Validity of peer nominations in predicting a distant performance criterion. *Journal of Applied Psychology, 49*(6), 434. https://doi.org/10.1037/h0022805

Krippendorf, K. (1970). Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement, 30,* 61-70. https://doi.org/10.1177/001316447003000105

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815-852. https://doi.org/10.1177/1094428106296642

Levi, M., Torrance, E. P., & Pletts, G. O. (1954). Sociometric studies of combat air crews in survival training. *Sociometry, 17*(4), 304-328. https://doi.org/10.2307/2785962

O'Donnell, A. M., & Topping, K. (1998). Peers assessing peers: Possibilities and problems. In K. Topping, & S. Ehly (Eds.), *Peer-Assisted Learning* (pp.255-278). Mahwah, NJ: Lawrence Erlbaum Associates.

Norton, S. M. (1992). Peer assessments of performance and ability: An exploratory meta-analysis of statistical artifacts and contextual moderators. *Journal of Business and Psychology, 6*(3), 387-399. https://doi.org/10.1007/BF01126773

Reynolds, H. H., (1966). Efficacy of sociometric ratings in predicting leadership success. *Psychological Reports, 19,* 35-40. https://doi.org/10.2466/pr0.1966.19.1.35

Schwarzwald, J., Koslowsky, M., & Mager-Bibi, T. (1999). Peer ratings versus peer nominations during training as predictors of actual performance criteria. *The Journal of Applied Behavioral Science, 35*(3), 360-372. https://doi.org/10.1177/0021886399353007

Sung, Y. T., Chang, K. E., Chang, T. H., & Yu, W. C. (2010). How many heads are better than one? The reliability and validity of teenagers' self- and peer assessments. *Journal of Adolescence, 33,* 135-145. https://doi.org/10.1016/j.adolescence.2009.04.004

U.S. Department of the Army. (2012). *Army Leadership,* (ADRP 6-22). Washington, D.C.: Author.

Wherry, R. J., & Fryer, D. H. (1949). Buddy Ratings: Popularity Contest or Leadership Criteria? *Personnel Psychology, 2*(2), 147-159. https://doi.org/10.1111/j.1744-6570.1949.tb01395.x

Williams, S. G., & Leavitt, H. J., (1947). Group opinion as a predictor of military leadership. *Journal of Consulting Psychology, 11(6),* 283-291. https://doi.org/10.1037/h0056512

Zazanis, M. M., Zaccaro, S. J., & Kilcullen, R. N. (2001). Identifying motivation and interpersonal performance using peer evaluations. *Military Psychology, 13*(2), 73. https://doi.org/10.1207/S15327876MP1302_01