Assessment Instruments in Support of Marine Instructor Development

Jennifer K. Phillips, Karol G. Ross
Cognitive Performance Group
Orlando, FL
jenni@cognitiveperformancegroup.com,
karol@cognitiveperformancegroup.com

Patrick J. Rosopa
Clemson University
Clemson, SC
prosopa@clemson.edu

ABSTRACT

Marine Corps instructors typically serve three-year assignments with no prior teaching experience. Although they may be subject-matter experts, the ability to pass knowledge to others is a distinct skill set and the training they receive to do so varies greatly. To maximize instructor performance, there is a need to accelerate the development of their teaching proficiency. To address this need, our team developed a mastery model for USMC instructors which describes the desired performance and provides a roadmap for development (Vogel-Walcutt, Phillips, Ross, & Knarr, I/ITSEC, 2015). The model was adopted as the basis for a new Training and Readiness Manual for instructors and as part of staff and faculty development policy. This paper reports on the next step, application of the model to develop and validate a Marine Instructor Assessment Toolkit. Assessment tools were developed to support the formal schools in accelerating instructor development with feedback-oriented instruments. They include an Observation Rubric for instructional settings, a Supervisor Rating Form for holistic instructor performance, a Self-Reflection Tool, and a Situational Judgment Test. The tablet- and web-based tools were field tested to gather user input at formal schools, and data were subjected to psychometric analysis which found the Observation Rubric and Supervisor Rating Form to be reliable and valid instruments. After field testing, the tools were finalized based on the psychometric analysis and user input. As the front end of transition to the formal schools, a baseline of instructor proficiency is currently in process to include a sample of up to 300 instructors. The baseline will serve as comparison data for future instructor performance measurements after full implementation of the instruments across the formal schools. Transition efforts following establishment of the baseline will consist of train-the-trainer workshops to familiarize the formal schools with the mastery model and the assessment tools.

ABOUT THE AUTHORS

Jennifer K. Phillips is the Chief Operating Officer and a Senior Scientist at the Cognitive Performance Group. Her research interests include skill acquisition, cognitive performance improvement, and the nature of expertise. Ms. Phillips applies cognitive task analysis and related techniques to model performance across the levels of proficiency, designs learning solutions including decision-centered training scenarios and facilitation techniques, and develops metrics for cognition and decision making. She is currently supporting Marine Corps programs to Accelerate the Development of Small Unit Decision Making, assess proficiency among Marine instructors, and support instructor development activities.

Karol G. Ross is the Chief Scientist for Cognitive Performance Group where she conducts applied research using qualitative and quantitative methods to develop performance models, training, and assessment in military environments. Her prior research includes a project to refine a general five-stage model of cognitive skills acquisition to support training development. She applied this model to the development of a model of tactical thinking and a Tactical Thinking Behaviorally Anchored Rating Scale for the US Army. Dr. Ross has been the PI for work in cross-cultural competence modeling and assessment. She recently served as the PI for an ONR research effort to develop a mastery model and an assessment battery to support USMC infantry squad leader development. She is currently the PI for the Master Instructor Development program (MInD) for ONR in support of the USMC which includes development of a mastery model and instructor assessment instruments. Dr. Ross earned a Ph.D. in Experimental Psychology from the University of Tennessee.

Patrick J. Rosopa is an Associate Professor in the Department of Psychology at Clemson University. His current research interests are related to individual differences, stereotypes, cross-cultural issues, research methods, and statistics. His research has been funded by $3.9 million in grants from the National Science Foundation and private industry. The results of Dr. Rosopa's research have been published in such peer-reviewed journals as *Human Resource Management Review*, *Personality and Individual Differences*, *Organizational Research Methods*, and *Psychological Methods*. He co-authored a statistics textbook through Wiley titled *Statistical Reasoning in the Behavioral Sciences* (6th ed.). In addition, Dr. Rosopa frequently serves as a consultant on research design and statistical analysis for various non-profit and for-profit organizations. He earned his B.S. in Psychology from Tulane University and his M.S. and Ph.D. in Industrial-Organizational Psychology from the University of Central Florida.

Assessment Instruments in Support of Marine Instructor Development

Jennifer K. Phillips, Karol G. Ross
Cognitive Performance Group
Orlando, FL
jenni@cognitiveperformancegroup.com,
karol@cognitiveperformancegroup.com

Patrick J. Rosopa
Clemson University
Clemson, SC
prosopa@clemson.edu

INTRODUCTION

Marine Corps active duty instructors serve in their teaching role for a mere three years, which for many individuals is an insufficient duration to naturally achieve mastery or even competency as a facilitator of learning. In response to the need to improve instructor professional development and the force effectiveness that high quality instruction engenders, the Marine Corps Training and Education Command (TECOM) is pursuing a faculty development initiative to further professionalize the instructor cadre throughout the Marine Corps. The specific objectives of the initiative are to reduce variation in instructor quality by institutionalizing a validated approach to instructor development; focusing on instructional techniques that facilitate higher levels of service-wide mastery and cognitive readiness; improving local-level support for instructor development; and raising the bar on instructor performance.

The challenge of raising the bar on instructor performance implies two important objectives. First, there is a need to define what is meant by “high quality instruction” and identify where the bar is to be set. Then, there is a requirement for a measurement capability to determine whether current or newly introduced instructor development activities are having the desired effect of improving teaching skills.

Current State of Instructor Development

Marine Corps train-the-trainer schools have traditionally offered courses to help prepare individuals for their instructional responsibilities, yet these courses are short in duration (i.e., five days each) and are reported to have limited effectiveness in building teaching skills. More recently, TECOM has produced a new Instructor Development Course (IDC) consisting of 30 days of distance coursework and eight days of resident work. Furthermore, many formal schools throughout the Marine Corps offer their own instructor development courses to address their unique domains and educational missions and extend the time and quality of the teaching skill instruction. Still, current instructor development practices are generally limited by practical constraints and a shortage of high quality supporting tools.

As a first step to raising the bar on instructor performance, the Office of Naval Research (ONR) sponsored a research effort in conjunction with TECOM to more clearly define instructor expertise and the path to mastery for Marine Corps instructors. The result of this effort was an Instructor Mastery Model, which is a developmental model highlighting the performance demands and requirements for instructor success (Ross, Phillips, & Lineberger, 2015; Vogel-Walcutt, Phillips, Ross, & Knarr, 2015). Specifically, the model produced ten Key Performance Areas (KPAs; see Table 1) as the central competencies of the formal teaching role for Marine instructors. Further, the model described the nature of performance along each KPA for individuals functioning at each of five stages of development—novice, advanced beginner, competent, proficient, and expert. The KPAs from the model were adopted as Marine Corps policy in the form of five training and readiness (T&R) events and five learning outcomes documented in the Train-the-Trainer T&R Manual (USMC, May 2015), and in the form of requirements for staff and faculty development plans in the Formal School Management Policy Guidance (USMC 1553.2, Sep 2015).

Table 1. Instructor Key Performance Areas

Key Performance Areas
1. Instructional Technique
2. Setting the Example
3. Communication and Delivery
4. Self-Improvement
5. Developing Subordinates and Peers
6. Planning and Preparation
7. Learning Environment
8. Assessing Effectiveness
9. Subject Matter Expertise
10. Community of Practice

The Marine Corps policy documents define a new standard for instructor performance. While the ultimate goal is to produce instructors who perform as experts in the teaching craft, the Marine Corps largely recognizes that few individuals will achieve expert level performance during the short three-year instructor tour. Instead, many schools have set the goal of producing competent, or Stage 3, individuals as the minimum level of performer to lead instructional sessions with students.

Objective

The purpose of this paper is to describe the development and validation of the Marine Instructor Assessment Toolkit (MIAT) to further support faculty development across the Marine Corps. ONR and TECOM identified the need to use the Instructor Mastery Model to produce an assessment capability to quantify current levels of performance across the instructor cadre, and to enable constructive feedback to be used in support of developing the desired levels of instructor performance. As a result of a review of current practices across the Marine Corps, we identified the following specific goals to guide the development of the MIAT:

1) Enable the formal schools to objectively measure performance, identify an instructor’s current stage of development within each of the ten KPAs, support individualized development plans based on relative strengths and development needs, and provide insights into the effectiveness of current instructor development practices;
2) Similarly, provide quantitative performance data at the Training Command, Education Command, or TECOM levels which would reveal trends and provide insights into the impact of Marine Corps-wide policies and interventions, enabling a continuous capability to improve upon faculty development practices;
3) Establish a common standard of instructional excellence across all schools by clearly describing each stage of instructor performance for the benefit of administrators of the tools and the instructors being evaluated;
4) Enable nuanced feedback regarding varying degrees of instructional effectiveness where current evaluation tools tend toward mastery/non-mastery assessments of whether a step was completed;
5) Support the schools in their compliance with the Formal School Management Policy Guidance by providing assessment tools specific to each KPA; and
6) Enable flexible application of the assessments across schools, recognizing the differences among the schools in their training versus education missions and instructional practices.

ASSESSMENT INSTRUMENT DEVELOPMENT

Initial development of the MIAT consisted of four steps. The first step was to understand the formal schools’ current processes for evaluating and developing instructors to ensure the Toolkit fits within the existing structures. In the second step, we identified candidate instrument types by deconstructing the performance defined within the Instructor Mastery Model and considering instruments amenable to use within the schools. Step three was to generate alpha versions of the instruments and their content. Finally, we conducted iterative subject-matter expert (SME) reviews and usability tests of the candidate instruments to continuously refine them. At the conclusion of these four steps, the instruments were subjected to field testing and psychometric analysis. This process yielded four instruments to meet the goals of the Marine Corps and specific measurement objectives derived from the Mastery Model:

1) A **Situational Judgment Test (SJT)** to objectively measure application of knowledge to real-life classroom situations and challenges.
2) An **Observation Rubric (OR)** as a rich and nuanced means by which an observer can evaluate an instructor’s performance during a period of instruction and provide better feedback than the current, procedurally-based forms used by many formal schools.
3) A **Supervisor Rating Form (SRF)** to address the instructor’s whole performance as a faculty member both inside and out of the learning environment.
4) A **Self-Reflection Tool (SRT)** to function as a self-assessment of performance across the KPAs.

We developed content for each of the instruments using the Mastery Model as a guide. Although the instruments are administered electronically via a tablet and web-based application, paper-and-pencil versions of the instruments were developed first. Phillips and Ross (2016) describe the development steps in detail.

Situational Judgment Test

Situational Judgment Tests are performance tests that present a dilemma in the form of a short vignette, or stem, and require test-takers to choose the best course of action or rate the goodness of alternative courses of action. They are often used to measure performance on cognitively complex tasks due to their ability to assess responses to nuanced situations while still producing quantifiable, psychometrically valid outcomes. Studies using meta-analysis have shown SJTs to have superior validity over traditional techniques for predicting job performance (McDaniel, Hartman & Grubb, 2003; McDaniel, Morgeson, Finnegan, Campion, & Braveman, 2001). Whereas observation of naturalistic task performance is subject to differences in observer scoring approaches and lack of standardization of the evaluation context across test-takers, SJTs mitigate these challenges. They are typically lower in internal consistency reliability than other instrument types because they are multidimensional in nature, addressing more than a single discrete construct as a tradeoff for achieving measurement of knowledge application in a pseudo-naturalistic context. However, they are effective in demonstrating learning as a result of training interventions and differentiating job performance.

To produce a maximum amount of data per item and increase the predictive potential of the SJT (e.g., Chan & Schmitt, 2002; McDaniel & Nguyen, 2001), we instructed participants to rate the effectiveness of each of five responses to the situation described in the stem. This is in contrast to other approaches where respondents select the best one response choice. The scoring approach employs a distance-from-expert calculation (e.g., Sacco, Schmidt, & Rogg, 2000; Wagner, 1987) where an expert model defines the most and least effective courses of action for each vignette.

The Instructor Mastery Model was consulted to generate response choices for each stem. All response choices were designed to align with the performance indicators in the model such that each response choice would be typical of instructors at a certain stage of development. Every item consisted of five response choices, but the response choices in a single item do not necessarily represent all five stages of development. We purposefully chose to assess whether participants could accurately evaluate the effectiveness of the response choices, even when the choices were generally ineffective or generally effective, or when two choices were of roughly equivalent effectiveness. In other words, high performers should recognize the quality of a course of action on its own merits rather than in comparison with the four other courses of action. In administration of the SJT via tablet, participants are shown one response choice at a time and instructed to rate it independent of the other choices.

Observation Rubric

The OR is a customized Behaviorally Anchored Rating Scale (BARS) for rating of performance of KPAs that are observable during a period of instruction. BARS have been traditionally employed in organizational settings to measure the effectiveness of individuals performing a wide range of tasks (Muchinsky, 2003). The *Code of Best Practices for Experimentation* for the Department of Defense (Albert & Hayes, 2002) promotes BARS as a means of conducting performance assessment without reliance on SMEs. BARS are a favored measurement technique because raters utilizing BARS are less prone to biases such as the halo effect of positive leniency (Muchinsky, 2003; Riggio, 2000). In developing the OR, we generated behavioral anchors based upon the Mastery Model performance descriptors for Stages 1 through 5 of the model. We included half-step ratings (i.e., 1.5, 2.5, etc.), making it a nine-point scale to enable raters to select a behavior more advanced than one stage yet not quite yet secure at the next stage. The OR is intended to be completed by a faculty member observing the lesson, as is currently done using evaluation checklists. We developed the OR using the measurement objectives defined for each KPA and subcategory, and especially the specific performance indicators drawn from the model and associated with each measurement objective for each stage. First, we identified measurement objectives that would be observable during a period of instruction. Next, we modified the performance indicators 1) for clarity and conciseness, and 2) to parse the indicator into a discrete behavior when multiple behaviors were represented within a single indicator. In this manner, behavioral indicators were distilled at each of the five levels of performance for the relevant measurement objectives. Next, we re-named the measurement objectives for ease of OR use by observers. As a result, discrete items were distinguished within each KPA and subcategory. In some cases, a measurement objective was divided into more than one item when the performance indicators were turned into discrete behavioral indicators.

Supervisor Rating Form

The SRF is also a BARS rubric identical in format to the OR. It is a customized rating of performance on a 9-point scale (1.0, 1.5, … 5.0) of each of the 10 KPAs, designed to be completed by the instructor’s direct supervisor(s). Since some of the KPAs are not directly observable during class periods (e.g., Community of Practice) or limited in their ability to be observed (e.g., Assessing Effectiveness), it was deemed necessary to generate another rubric whereby human judgment could be used to produce a quantifiable assessment of performance. Whereas the OR focuses on what is observed during a single classroom session, the SRF is intended to comprise the performance observed in and out of the classroom over an extended period of time, such as annually or semiannually. We developed the SRF by following the same process as described above for the OR, however, every KPA was included and every measurement objective observable by a supervisor was addressed.

Self-Reflection Tool

The SRT is not an assessment instrument, rather it is designed to facilitate instructor development via reflection and self-assessment. The SRT is a customized self-rating of performance of each of the 10 KPAs using the content of the SRF, whereby the instructor rates his or her performance on a 9-point scale on each of the 10 KPAs. In addition, and more importantly, the instructor identifies development goals associated with each KPA. The SRT is intended to be completed annually or semiannually by the instructor in conjunction with the supervisor's completion of the SRF. The two sets of ratings are then to be compared and discussed in a coaching meeting.

FIELD TEST METHOD

In the field test, we collected data to validate three of the four instructor assessment tools—OR, SRF, and SJT—with a population of formal school instructors. The SRT was not tested because it serves as a self-assessment tool where the data output does not contribute to the performance score of the instructor.

Participants

A total of 125 instructors from 14 formal schools on and around Camp Pendleton, CA, Twentynine Palms, CA, Camp Lejeune, NC, and Quantico, VA, and representing Training Command (*N*=104), Education Command (*N*=6) and the Marine Air-Ground Task Force Training Center (MAGTF-TC; *N*=15), participated in the field test. Participants ranged from 0-244 months serving as instructors (see Figure 1); the mean time as an instructor was two years (*M*=24.4 months, *SD*=34.2 months). Those instructors' supervisors as well as individuals who serve regularly as classroom observers at the school also participated (*N*=89) by providing the SRF and OR data, respectively, regarding the instructor participants' performance. Note that a supervisor or an instructor participant could serve in a dual capacity as an observer in this study. Data about the supervisors and observers were not subject to analysis except in the inter-rater reliability calculations.

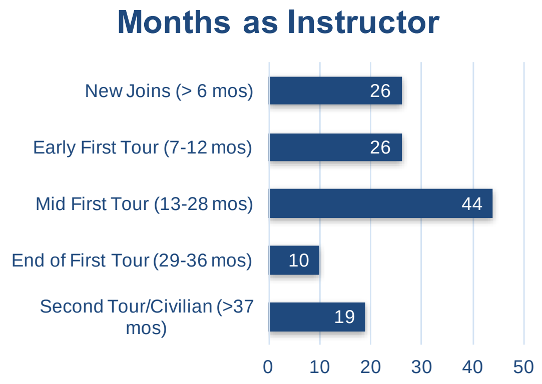

Figure 1. Range of participant time as instructor.

Materials

Participants completed a paper-based informed consent form and demographics form prior to administration of the assessment instruments. The three assessment instruments undergoing testing were administered in electronic form. The SRF was completed via the web-based application. The OR and SJT were completed via tablet application.

Procedure

The research team conducted a five-day data collection effort at each of the four geographical sites. Data collected included four components for each instructor participant: 1) a demographic form; 2) a SRF completed by an actual supervisor or acceptable proxy; 3) an OR completed by a qualified observer following a period of instruction led by the instructor participant; and 4) at least one of the two forms of the SJT (some instructors completed both Forms A and B to enable assessment of the form equivalency).

Inter-rater reliability of the two rating rubrics (i.e., the SRF and the OR) was assessed at each geographic site. Approximately 12% of the dataset, or at least 16 instructors, were initially targeted for assessment of inter-rater reliability; however, in practice and due to scheduling constraints, only 12 instructors had multiple raters apply the SRF to assess them. For the OR, 26 instructors had at least two raters contributing to the inter-rater reliability assessment. At every site, we identified rater pairs to contribute to the assessment of inter-rater reliability. After their independent ratings, we immediately examined their levels of agreement by item. We then assembled a meeting with the two raters and at least one researcher to discuss rater disagreements. For the items on which ratings differed by 1.5 point or more (e.g., Observer 1 rated a 4.0 and Observer 2 rated a 2.0), the discussion sought to reach concurrence and raters were given the opportunity to revise their ratings. In addition, those low-agreement items were noted so the training session for the next field test session could be adjusted to better calibrate raters on the meaning of the descriptors within the item.

PSYCHOMETRIC ANALYSIS FINDINGS

We conducted three phases of psychometric analysis of the field test data to assess the MIAT and the individual instruments it contained: *Instrument Internal Analysis, Instrument to Instrument Relationship Analysis,* and *Instrument and Whole Test to Criterion Analysis*. Each phase and its findings are described below.

Phase 1: Instrument Internal Analysis

In the *Instrument Internal Analysis* phase, the objective was to identify whether the items are good contributors to each individual instrument and whether the instruments are of acceptable quality. Specifically, we conducted an item analysis to determine items that did not perform well. We assessed the inter-rater reliability of the two rating rubrics. We examined whether an SJT expert model could be derived from the data set, and we assessed whether the two forms of the SJT were equivalent.

Observation Rubric Reliability

The field test version of the OR consisted of 29 items measuring five KPAs. To obtain an overall score on the OR, a composite of the five KPAs was also calculated. Based on user feedback during field testing and the results of an item analysis, the 29 items were reduced to 24 items. Internal reliability estimates calculated for the KPA variables once items were removed that did not contribute to reliability are presented in Table 2. Based on these results involving the OR, the proportion of the total variance within a KPA that is due to systematic variability was generally high and ranged between .802 and .917. The estimated reliability of the composite was also high, .959. We conclude that the OR items within each KPA are highly related, as expected because they were derived from the Mastery Model.

Table 2. Reliability Estimates for Observation Rubric

KPA Score Variable	*M*	*SD*	*N*	# of Items	Reliability Estimate
Instructional Technique	3.26	.94	119	9	0.917
Communication & Delivery	3.53	.87	121	5	0.912
Learning Environment	3.50	.84	121	4	0.848
Assessing Effectiveness	3.30	.90	119	3	0.858
Subject Matter Expertise	3.68	.89	121	3	0.802
Composite	3.44	.74	118	5	0.959

Inter-rater reliability was also calculated for all KPAs measured in the OR. Inter-rater reliability estimates ranged between .47 and .56. Although these values are less than .70, this range of values is consistent with extant research. For example, according to Schmidt, Viswesvaran, and Ones (2000), the average inter-rater reliability of ratings of job performance has been found to be .50.

Supervisor Rating Form Reliability

The field test version of the SRF consisted of 54 items that measure 10 KPAs. A composite score was also computed using the scores on the 10 KPAs. Based on user feedback during field testing and the results of an item analysis, the 54 items were reduced to 41 items. Estimated internal consistency reliabilities for each KPA assessed in the SRF, once items that did not contribute were removed, are presented in Table 3. Overall, reliability estimates for the individual

KPAs were high ranging from .822 to .952. The reliability estimate for the composite cannot be computed because we do not have scores on the Community of Practice KPA to form a composite of all 10 KPAs. As with the OR, we conclude that the SRF items within each of the first nine KPAs are highly related.

Inter-rater reliability was also calculated for all KPAs measured in the SRF. Reliability estimates ranged between .821 and .946. This exceeds conventional thresholds, suggesting that the ratings are reliable.

Table 3. Reliability Estimates for Supervisor Rating Form

KPA Score Variable	*M*	*SD*	*N*	# of Items	Reliability Estimate
Instructional Technique	3.34	.79	81	4	0.952
Setting the Example	3.60	.73	82	4	0.901
Communication & Delivery	3.46	.68	83	4	0.914
Self-Improvement	3.14	.79	64	5	0.941
Developing Subordinates & Peers	3.46	.79	83	5	0.941
Planning & Preparation	3.24	.84	81	6	0.951
Learning Environment	3.49	.71	82	5	0.939
Assessing Effectiveness	3.27	.78	82	4	0.947
Subject Matter Expertise	3.62	.73	81	2	0.822
Community of Practice			--	2	--
Composite			--	10	--

Situational Judgment Test Equivalency

Two forms of the SJT were administered—Form A and Form B—each consisting of 20 items. To score the SJT, an expert model was derived for each response choice on each item (for a detailed description of the expert model development, see Ross, Rosopa, & Phillips, 2017). An expert was defined as an individual who was identified as being in the top 10% on the same KPAs in the OR and SRF that were assessed in the SJT. Participant SJT scores were then calculated based upon their similarity to the expert model where a score of 1 means perfect agreement with the expert model, and a score closer to 0 means no agreement with the experts.

The correlations of KPA scores on Form A and Form B were computed to assess the equivalency of the forms. See the last column in Table 4. For example, the Instructional Technique KPA score in Form A was positively correlated with the Instructional Technique score in Form B (*r* = .47, *p* < .001), suggesting that a higher score on Instructional Technique in Form A was associated with a higher score on that KPA in Form B. Although the sample size for these correlations was 37, all correlations were positive and three out of five correlations were statistically significant at the .001 level. It deserves noting that although the mean differences are not substantial between Form A and Form B, given the positive correlations and small standard deviations, the paired samples *t* tests on the KPAs were all statistically significant at the .05 level. Thus, on Instructional Technique, for example, the mean difference of .41 and .44 between Form A and Form B, respectively, was statistically significant. From a practical perspective, this mean difference is not considered large, suggesting the equivalence of the two forms.

Table 4. Means and Standard Deviations on the KPAs of the SJT on Forms A and B

	Form A		Form B		
	M	*SD*	*M*	*SD*	*r*
Instructional Technique	0.41	0.06	0.44	0.07	0.47**
Self-Improvement	0.42	0.09	0.45	0.08	0.56**
Planning & Preparation	0.46	0.07	0.43	0.09	0.15
Learning Environment	0.40	0.06	0.46	0.07	0.28
Assessing Effectiveness	0.45	0.07	0.42	0.06	0.40**

Note. *N* = 37. ** *p* < .001.

Phase 2: Instrument to Instrument Relationships

In the *Instrument to Instrument Relationship Analysis* phase, the goal was to identify whether the instruments are good contributors to measuring KPA performance once problematic items identified in Phase 1 are removed. Specifically, we hypothesized different instruments would produce similar KPA scores, for example, the Instructional Technique KPA score on the OR would be positively correlated with the same KPA score produced by the SRF and the SJT.

Observation Rubric and Supervisor Rating Form Correlations

Instructional Technique, Communication & Delivery, Learning Environment, Assessing Effectiveness, and Subject Matter Expertise were assessed in both the OR and SRF; therefore, we computed correlations on these five KPAs. Table 5 presents the correlations on these five KPAs between the two instruments. There was clear evidence that the estimated correlations were positive, suggesting both rubrics measure similar competencies, and performance on each of these KPAs is related to performance on the other KPAs.

Table 5. Correlations Between KPAs on the Observation Rubric and the Supervisor Rating Form

	Supervisor Rating Form				
Observation Rubric	Instructional Technique	Communication & Delivery	Learning Environment	Assessing Effectiveness	Subject Matter Expertise
Instructional Technique	**0.583**	0.473	0.519	0.521	0.477
Communication & Delivery	0.615	**0.567**	0.526	0.516	0.467
Learning Environment	0.549	0.489	**0.484**	0.476	0.468
Assessing Effectiveness	0.585	0.478	0.482	**0.488**	0.461
Subject Matter Expertise	0.477	0.390	0.407	0.398	**0.469**

Note. N varies from 69 to 73. All correlations significant at *p* < .001.

Observation Rubric and Situational Judgment Test Correlations

Instructional Technique, Learning Environment, and Assessing Effectiveness were assessed in both the OR and SJT; therefore, we computed correlations on these three KPAs. The correlations were not significant between the OR and either form of the SJT.

Supervisor Rating Form and Situational Judgment Test

Instructional Technique, Self-Improvement, Planning & Preparation, Learning Environment, and Assessing Effectiveness were assessed in both the SRF and SJT; therefore, we computed correlations on these five KPAs. Table 6 presents these correlations for both forms of the SJT. Although some of the Form A correlation coefficients were positive, only a few were statistically significant and none of the correlations were significant for matching KPAs. In Form B, the number of complete observations was not large, but all the correlation coefficients were positive, with a few statistically significant. In general, higher scores on the KPAs in the SRF, the higher the scores on Form B of the SJT. However, matching KPAs from the two instruments were not significantly correlated.

Table 6. Correlations Between KPAs on the Supervisor Rating Form and the Situational Judgment Tests

	Situational Judgment Test, Form A					Situational Judgment Test, Form B				
Supervisor Rating Form	Instruct Tech	Self-Improve	Plan & Prep	Learn Env	Assess Effect	Instruct Tech	Self-Improve	Plan & Prep	Learn Env	Assess Effect
Instructional Technique	**0.003**	0.142	0.003	0.263*	0.010	**0.356**	0.379*	0.414*	0.324	0.244
Self-Improvement	0.095	**0.246**	0.010	0.185	0.088	0.348	**0.196**	0.416*	0.368	0.266
Planning & Preparation	0.023	0.194	**0.108**	0.295*	0.172	0.304	0.325	**0.333**	0.143	0.127
Learning Environment	0.037	0.237*	0.193	**0.428**	0.177	0.319	0.381	0.494*	**0.299**	0.298
Assessing Effectiveness	0.038	0.229*	0.151	0.438*	**0.201**	0.397*	0.436*	0.513*	0.359	**0.310**

Note. For Form A, *N* varies from 58 to 78. For Form B, *N* varies from 27 to 29. * *p* < .05.

Phase 3: Instrument and Whole Test to Criterion Analysis

In the third phase, *Instrument and Whole Test to Criterion Analysis*, the purpose was to determine whether KPA scores are meaningful and useful in support of instructor development. We expected the KPA scores to be predictive of a global assessment criterion rating. We expected time in an instructor billet to be positively correlated with KPA scores but time in service to not show the same correlation, suggesting that better teachers require more time teaching and not necessarily more time as a Marine. Finally, we expected to find a positive relationship between KPA scores and the amount of instructor-specific training and preparation provided.

KPA Scores and Global Criterion

The KPA scores from the OR, SRF, and SJT were used to predict the global criterion rating, resulting in three multiple regression analyses. In the first multiple regression analysis using the five KPAs represented on the OR to predict the global criterion, the overall model was statistically significant, F (5, 79) = 5.521 (p < .001), explaining 25.9% of the variance in the global criterion ratings. However, none of the regression coefficients were statistically significant. An inspection of the variance inflation factors associated with each term confirmed that the values ranged between 3.9 and 9.5, suggesting strong relationships among the five KPAs was likely impacting the analysis. These findings suggest the OR as a whole provides a meaningful assessment of the stage of proficiency of an instructor, but it is unclear which specific KPAs contribute towards an instructor's overall proficiency.

In the second multiple regression analysis using the KPAs on the SRF to predict the global criterion, the overall model was statistically significant, F (9, 48) = 18.517 (p < .001), explaining 77.6% of the variance in the global criterion ratings. The results of the second multiple regression analysis, including the regression coefficients, standard errors, and t statistics, are presented in Table 7. Instructional Technique, Communication and Delivery, Developing Subordinates and Peers, Planning & Preparation, Learning Environment, and Subject Matter Expertise were statistically significant. This finding indicates the SRF provides a more valid overall assessment of instructor proficiency than the OR, which is to be expected considering the SRF addresses all 10 KPAs. Further, six of the 10 KPAs may be of greater importance to overall instructor proficiency than the other four KPAs.

Table 7. Multiple Regression Analysis Predicting Global Criterion Using KPAs on Supervisor Rating Form

	B	*SE*	*β*	*t*	*p*
Intercept	0.34	0.30		1.13	0.265
Instructional Technique	0.46	0.18	0.48	2.53	0.015
Setting the Example	-0.23	0.17	-0.21	-1.34	0.188
Communication and Delivery	0.54	0.23	0.47	2.34	0.024
Self-Improvement	0.08	0.17	0.08	0.48	0.636
Developing Subordinates and Peers	-0.51	0.17	-0.50	-3.07	0.003
Planning & Preparation	0.37	0.17	0.38	2.25	0.029
Learning Environment	0.52	0.23	0.45	2.28	0.027
Assessing Effectiveness	-0.09	0.17	-0.08	-0.51	0.613
Subject Matter Expertise	-0.27	0.13	-0.27	-2.17	0.035

Note. N = 58. Multiple R = .881.

In the third multiple regression analysis, whether using the KPAs from the SJT Form A or Form B to predict the global criterion, the overall model was not statistically significant. However, because of the larger sample size for Form A (N = 87) compared to Form B (N = 26), the multiple regression involving Form A approached statistical significance (p = .076), explaining 11.4% of the variance in the global criterion ratings.

Because the KPAs from the OR and SRF were statistically significant in predicting global criterion ratings, additional regression analyses were conducting using the KPA composites to predict global criterion ratings. Both simple linear regression analyses were statistically significant. The prediction equations for each and the squared multiple correlation are presented below.

$$\hat{y}_{Criterion} = 1.43 + 0.54(OR) \qquad R^2 = .202 \qquad (1)$$

$$\hat{y}_{Criterion} = -0.06 + 0.99(SRF) \qquad R^2 = .631 \qquad (2)$$

Thus, the OR accounted for 20.2% of the variance in the global criterion. The SRF accounted for 63.1% of the variance in the global criterion.

KPA Scores, Time as Instructor, and Time in Service

Pearson product-moment correlation coefficients were calculated between KPA scores and time as an instructor, and between KPA scores and time in service. The correlation coefficients between KPA scores from the OR and time as an instructor were all positive, but none were statistically significant. The correlation coefficients between KPA scores from the SRF and time as an instructor, and from the SJT and time as an instructor were likewise not statistically significant. Similarly, none of the KPA scores for the OR, SRF, or SJT were related to time in service. Taken together, neither time as instructor or time in service were related to KPA scores.

Previous Formal Training and KPA Scores

To provide evidence that the KPA scores can discriminate between performance based on previous formal training, mean differences were examined. Because of the limited sample sizes across groups, two groups were formed to represent less versus more formal training. For the OR, there were statistically significant mean differences on Instructional Technique, Communication & Delivery, Learning Environment, and Subject Matter Expertise (see Table 8). Thus, respondents with more formal training tended to have higher average scores on these four KPAs compared to those with less formal training.

Table 8. Two Independent Sample *t* tests on Observation Rubric KPA Scores using Formal Training as a Grouping Variable

	Low			High			
Variable	*N*	*M*	*SD*	*N*	*M*	*SD*	*t*
Instructional Technique	66	3.11	0.72	39	3.44	0.72	2.31*
Communication & Delivery	67	3.32	0.69	39	3.84	0.73	3.65**
Learning Environment	67	3.37	0.64	39	3.65	0.72	2.04*
Assessing Effectiveness	66	3.23	0.79	38	3.36	0.79	0.80
Subject Matter Expertise	67	3.52	0.76	39	3.83	0.71	2.03*
KPA Composite	66	3.31	0.66	38	3.61	0.69	2.22*

Note. Low = less formal training. High = more formal training.
* $p < .05$. ** $p < .01$.

For the SRF, there were statistically significant mean differences on all nine KPAs (see Table 9). Thus, respondents with more formal training tend to have higher average scores on all nine KPAs compared to those with less formal training.

For the SJT, there were no statistically significant mean differences on the KPAs from Form A. On Form B, there was a statistically significant mean difference ($p < .034$) on Instructional Technique such that those with more formal training ($M = 0.47$, $SD = 0.05$) tended to have higher average scores than those with less formal training ($M = 0.42$, $SD = 0.08$).

Table 9. Two Independent Sample *t* tests on Supervisor Rating Form KPA Scores using Formal Training as a Grouping Variable

	Low			High			
Variable	*N*	*M*	*SD*	*N*	*M*	*SD*	*t*
Instructional Technique	55	3.18	0.78	26	3.68	0.72	2.76**
Setting the Example	55	3.46	0.67	28	3.87	0.79	2.51*
Communication & Delivery	55	3.29	0.64	28	3.81	0.63	3.57**
Self-Improvement	43	2.93	0.76	21	3.57	0.69	3.24**
Developing Subordinates & Peers	55	3.28	0.74	28	3.84	0.78	3.23**
Planning & Preparation	54	3.04	0.78	27	3.65	0.85	3.25**
Learning Environment	55	3.32	0.61	28	3.83	0.76	3.32**
Assessing Effectiveness	54	3.09	0.73	28	3.63	0.75	3.18**
Subject Matter Expertise	53	3.49	0.73	28	3.87	0.83	2.13*
KPA Composite	56	3.26	0.64	28	3.78	0.71	3.41**

Note. Low = less formal training. High = more formal training.
* $p < .05$. ** $p < .01$.

DISCUSSION AND CONCLUSIONS

The purpose of developing a validated MIAT was to support Marine Corps efforts to accelerate the development of instructor proficiency. At the formal school level, the MIAT was intended to provide nuanced assessment and feedback tools. At the TECOM and major subordinate command levels, the goal was to provide objective and quantitative assessments of instructor proficiency to produce trend data sufficient for analyzing the impact of policies and interventions on instructor skills across the force.

The findings of the psychometric analysis suggest the OR and SRF to be reliable and valid instruments for assessing instructor proficiency. They both show high internal consistency reliability within each KPA and across the KPAs collectively. Although the inter-rater reliability of the OR is lower than desired, we believe additional training and calibration in the tool's use at each schoolhouse may increase rater agreement to sufficient levels. To that end, we have developed a video-based training tool to demonstrate each of the behaviors assessed in the OR and will re-assess inter-rater reliability with the use of this pre-observation training. In addition, the KPA scores produced by the OR and SRF are correlated across instruments, and the composite scores on the OR and SRF are correlated with the global criterion variable suggesting criterion validity. It is notable that the 10 KPAs appear to be correlated with each other as much as with the matching KPA score from the other rating rubric. We conclude from this finding that the KPAs are not independent constructs yet all contribute to the proficiency of an instructor. Additional factor analyses must be conducted to determine whether the rubric items cluster more meaningfully into a set of constructs other than the 10 KPAs defined by the Mastery Model. Furthermore, the finding that six of the ten SRF KPA scores correlate significantly with the criterion variable indicates the need for additional research to determine whether those KPAs are more critical to instructor proficiency, or whether another interpretation accounts for the finding.

Field test participants reported they found the OR and SRF to be useful for providing feedback to instructors. Specifically, they believe the ability to establish a common language for describing the elements of instructor performance contributing to student learning, as well as the nature of distinct instructor skill levels, are of great value to their instructor development efforts. By and large, the schools favor the rubrics as tools to facilitate discussions and qualitative performance feedback as opposed to quantitative assessment scores.

While the SJT analyses suggest the two forms to be equivalent in their measurement of instructors, the other analysis outcomes do not suggest the SJTs are measuring KPA skills as intended. Form B demonstrated positive but not significant correlations between matching KPA scores on the SRF and SJT; Form A showed positive but smaller and insignificant correlations. However, the existence of correlations among some of the non-matching KPAs between the SRF and SJT suggest the SJT has value in assessing instructor proficiency. Additional research and analysis are required to determine whether the SJT has utility for assessing instructors. To that end, we plan to: (a) re-assess the process for defining expert responses in the expert model (e.g., using an independent sample of experts rate the response choices for each item), (b) examine alternative scoring approaches (e.g., utilizing different distance measures to compute a similarity score), and (c) review the forms at the item level to determine whether removal of individual items or response choices improves the instrument's measurement ability.

The MIAT in its current form demonstrates value, and marked improvement over current approaches, for supporting the formal school goal of providing nuanced feedback to instructors to support their skill development, especially with the use of the OR and SRF. The next step is to initiate implementation of these tools into the formal schools to collect additional data to address the follow-on research questions, and to more clearly define and support users' needs related to integration of the tools into their instructor development practices. For example, what are the best practices for calibrating raters' application of the rating rubrics, and are instructors better served by receiving feedback about and working on improving all KPAs simultaneously or a few at a time? To that end, we are in the process of conducting train-the-trainer workshops with school personnel to hand off the instruments for collection of data by the schools instead of the research team. Our team will analyze the resulting data and provide results to the schools and to TECOM as a baseline measurement of instructor proficiency, and user feedback will be collected to inform refinement of the assessment instruments, supporting tools, and future development of activities to improve instructor skills.

Significant progress has also been made toward the TECOM-level goal of collecting performance trends among instructors using objective and quantitative assessments administered across the schools. Although additional testing is required to modify and establish the validity of the SJT, it demonstrates potential as one such measure of proficiency that can be administered without the involvement of raters, and at regular (e.g., annual) intervals during an instructor's tenure. In addition, both the OR and the SRF are quantitative performance measures whose objectivity can be established with additional structured training for inter-rater reliability administered across all participating schools. The next step to achieve TECOM's goal is to determine the mechanism by which data collected at the local school level can be viewed and analyzed at the TECOM level without personally identifying the individuals associated with the data or imposing onerous data transmission requirements upon the schools.

With the generation of the Instructor Mastery Model to define the new standard for Marine Corps instructors and the development of the MIAT which provides an improved instructor assessment capability, the Marine Corps has achieved two significant outcomes in its quest to improve the quality of instruction across the force. Future research and development must identify and/or develop specific activities and tools that accelerate the development of instructor proficiency. The MIAT instruments provide a description of an individual's current stage of development along each KPA. Interventions for enhancing instructor skill, therefore, may target individual development needs by providing activities that will move instructors from their current level of proficiency to the next.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Peter Squire of ONR and his team for their sponsorship, insight, and close collaboration. Colonel Michael Watkins, Major Nicholas Armendariz, Major Eli Jones, Mr. Kenn Knarr, Mr. John DeForest, and Mr. William Morris were instrumental in their support and guidance. Finally, many thanks to the schools across the Marine Corps for their time and valuable feedback.

REFERENCES

Alberts, D.S., & Hayes, R.E. (2002). *Code of best practices for experimentation*. Retrieved from http://www.dodccrp.org/publications/pdf/Alberts_Experimentations.pdf.

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A., & Braveman, E.P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730-740.

McDaniel, M.A., Hartman, N.S. & Grubb III, W.L. (2003, April). *Situational Judgment Tests, Knowledge, Behavioral Tendency, and Validity: A Meta-Analysis*. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology. Orlando.

Muchinsky, P.M. (2003). *Psychology applied to work*. Melmont, CA: Wadsworth/Thomson Learning.

Phillips, J.K., & Ross, K.G. (2016). Option III: Instructor assessment battery development technical report, part 2. Technical report produced under Contract N00014-14-C-0106 for the Office of Naval Research. Orlando, FL: Cognitive Performance Group.

Riggio, R.E. (2000). *Introduction to industrial/organizational psychology*. Upper Saddle River, NJ: Prentice Hall.

Ross, K.G., Phillips, J.K., & Lineberger, R.E. (2015). Marine Corps Instructor Mastery Model. Technical report produced under Contract N00014-14-C-0106 for the Office of Naval Research. Orlando, FL: Cognitive Performance Group.

Ross, K.G., Rosopa, P.J., & Phillips, J.K. (2017). Marine Corps Instructor Assessment Toolkit Psychometric Analysis. Manuscript in process.

Sacco, J.M., Schmidt, D.B., & Rogg, K.L. (2000, April). *Understanding race differences on situational judgment tests using readability statistics*. Paper presented at the 15th Annual Convention of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Schmidt, F.L., Viswesvaran, C., & Ones, D.S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*, 901-912.

U.S. Marine Corps (1 May 2015). NAVMC 3500.37C, Train the Trainer Training and Readiness Manual. Washington, DC: Department of the Navy.

U.S. Marine Corps (21 Sep 2015). NAVMC 1553.2, Marine Corps Formal School Management Policy Guidance. Washington, DC: Department of the Navy.

Vogel-Walcutt, J.J., Phillips, J.K., Ross, K.G., & Knarr, K.A. (2015). Marine Corps instructor mastery model: A foundation for Marine faculty professional development. Proceedings of the Interservice/Industry Training, Simulation, and Education Conference. Orlando, FL: NDIA.

Wagner, R.K., (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology, 52*, 1236-1247.