Wright State University
CORE Scholar

International Symposium on Aviation Psychology - 2005 | International Symposium on Aviation Psychology

2005

Evidence Against Crew Resource Management as a Cognitive Skill

Stacey M. L. Hendrickson

David L. Trumpower

Timothy E. Goldsmith

Peder J. Johnson

Follow this and additional works at: https://corescholar.libraries.wright.edu/isap_2005

 Part of the Other Psychiatry and Psychology Commons

Repository Citation

Hendrickson, S. M., Trumpower, D. L., Goldsmith, T. E., & Johnson, P. J. (2005). Evidence Against Crew Resource Management as a Cognitive Skill. *2005 International Symposium on Aviation Psychology*, 299-303.
https://corescholar.libraries.wright.edu/isap_2005/43

This Article is brought to you for free and open access by the International Symposium on Aviation Psychology at CORE Scholar. It has been accepted for inclusion in International Symposium on Aviation Psychology - 2005 by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu, library-corescholar@wright.edu.

EVIDENCE AGAINST CREW RESOURCE MANAGEMENT AS A COGNITIVE SKILL

Stacey M. L. Hendrickson
University of New Mexico
Albuquerque, New Mexico

David L. Trumpower
Marshall University
Huntington, West Virginia

Timothy E. Goldsmith
University of New Mexico
Albuquerque, New Mexico

Peder J. Johnson
University of New Mexico
Albuquerque, New Mexico

In recent years, airlines have begun to train and assess crew resource management (CRM) tasks similarly to technical tasks. However, in order for individual CRM categories (e.g., workload management, communication, situation awareness, etc.) to be viewed as skills, performance on a particular CRM category should transfer to different situations. In this study, we examined how well CRM behaviors generalized across different flight contexts. We analyzed pilot performance from five line oriented evaluations (LOEs). The LOEs were divided into phases of flight and many different behaviors were graded within each LOE, some of which were previously classified as belonging to a particular CRM category (e.g., workload management). A series of regression analyses showed that less than 1% of the total variance in grades was due to CRM categories; in contrast phase of flight accounted for roughly 10% of the total variance in grades. Thus, pilots performed more consistently within a phase of flight (regardless of CRM task category) than within a specific CRM category. We discuss several caveats and limitations associated with these findings. However, the findings do question the idea that CRM performance is a skill. One implication of these results is that pilot training may be more effectively focused around contexts rather than around specific CRM task categories.

Introduction

According to Welford (1976), a skill is defined by a high level of performance on a task that is achieved through training, is relatively permanent, and generalizes across similar situations. Training that aims to foster skill acquisition assumes that skills will generalize to contexts outside of training. Clearly this assumption is warranted in a broad sense; students do indeed gain expertise and go on to perform well in novel, real-world situations. However, questions remain about what kinds of task performances are best viewed as skills and what are the best methods for training and assessing these performances. These questions appear to be of particular relevance to the aviation community with respect to crew resource management (CRM). In recent years airlines have begun to train and assess CRM task performance in much the same way as technical skills. In this study, we investigated to what extent performance on CRM tasks such as situation awareness, decision-making, and workload management could be viewed as skills.

A related question is how effective is CRM training. In a recent review of its effectiveness, Salas et al. (in press) found mixed results, particularly when the assessment of CRM effectiveness was focused on behavioral outcomes. Other researchers have questioned CRM's effectiveness and validity too (Wiener, Kanki, & Helmreich, 1993). Perhaps CRM training has not been found to be more effective because the knowledge and behaviors being trained are in fact not skills, at least not in the traditional sense. An alternative and competing viewpoint is that CRM performance is contextually specific; performance on CRM-type tasks is more of a function of specific flight situations than the category of CRM behavior. If true, this would have important implications for training and assessing CRM.

Rather than examining the effectiveness of CRM training per se, we sought evidence that CRM performance behaved as an enduring trait across varied situations. More specifically, we investigated the construct validity of CRM categories by examining to what extent a pilot's performance in a single CRM category (e.g., decision making) was similar across

two or more samples of this performance. If we know how a pilot performs on a particular CRM skill in one context, then we should be able to predict his performance for that same type of task in a different context. Conversely, the categories should also display discriminant validity in that we should be able to discriminate between the crew's performance on different CRM categories. We applied the classic psychometric paradigm of multi-trait multi-method (Campbell & Fiske, 1959) for investigating traits and situations to assess CRM skills of pilots. In the current study, we analyzed performance from a set of pilots who were evaluated along various CRM tasks that occurred in different contexts, in this case, phases of flight. If performance on CRM task categories is indeed skill-like, then we would expect to find higher similarity in performance of pilots within a CRM category than within a context.

Method

The basic data set we analyzed consisted of pilot performance data from 348 crews performing five LOEs under continuing qualification. Each crew was evaluated by an instructor/evaluator (IE) on the performance of 72 observable behaviors (OBs) over the course of an entire flight. Each LOE was comprised of 12 event sets (ES), each associated with a phase of flight (e.g., take-off, cruise, landing). Each crew was assessed on a 5-point grade scale by one of 20 IEs in one of the five LOEs.

Thirty-five of the OBs were intended to measure CRM performance (divided into five categories), and the other 37 OBs measured technical skills (divided into four categories). CRM categories were represented with OBs that might focus on the crew's communication (e.g. "The crewmembers state their ideas, opinions, and/or recommendations") or, as another example, they might address the crew's situational awareness (e.g. "The crew maintains shared level of situational awareness during precision approach").

Results

The grade distribution for all grades received by the crews are displayed in Table 1. Also shown is the breakdown for the grade distribution for CRM OBs and technical OBs. In all instances, the grade distribution was skewed so that the grades given were a majority of passing (greater than 2).

Table 1. *Proportion of grades received by flight crews*

	1 (unsatisfactory)	2	3	4	5 (excellent)	Mean (Std)
All OBs	<1	2	20	49	28	4.03 (.74)
CRM OBs	<1	1	24	47	28	4.02 (.74)
Technical OBs	<1	2	20	50	28	4.05 (.74)

Several multiple regression analyses were performed in order to assess the influence of the factors of specific CRM skill, context, and general skill on OB grades. Separate analyses were performed on CRM and technical grades. The dependent variable for each of these regression analyses was a crews' OB grade. Thus, each case corresponded to a single crew's grade on a single OB. For the CRM analysis, any given crew was represented by 35 cases since there were 35 CRM OBs in an LOE. With data from 348 crews, there were over 12,000 cases.

In the first analysis, we created predictor variables that reflected a crew's general skill, their performance on a particular skill category, and their performance in a given context. The three predictor variables were constructed as follows. First, for a given case, a mean grade was calculated for that particular crew's grades on all OBs from *different ESs* and *different skill* categories. This predictor was called General Skill (GS) as it represented a crew's performance across many phases of flights and skills. Second, an average was calculated for the particular crew's grades on all OBs of the *same skill* and *different ESs*. This predictor reflects a crew's performance in a particular CRM skill, and is thus called Skill (S). Third, an average was calculated for a crew's grades on all OBs from *different skills* but from the *same ES*. As this predictor represents a crew's performance in a specific event set, it was deemed Context (C).

For an example of how these variables were computed, consider a case associated with crew number 1 and a CRM-1 OB in the first event set. The GS predictor score for this case would be the average of crew number 1's grades on all non-CRM-1 CRM OBs in the second through twelfth ESs, the S predictor score would be the average of crew number 1's grades on all CRM-1 OBs in the second through twelfth ESs, and the C score would be the average of that crew's grades on all non-CRM-1 OBs in the first ES. Notice that all three predictors are orthogonal to one another; none of a crew's grades are used in the computation of more than one of the predictors.

From the regression analyses, a table can be constructed of the simple and semi-partial correlations among the CRM grades and the set of predictor variables. Results of this analysis are shown in Table 2. Overall, the three predictors accounted for 53.5% of the variability in CRM OB ra-tings. Although all three of the effects contributed significantly to prediction, the context effect was the strongest predictor accounting for 9.8% unique variance, while the skill and general skill effects uniquely accounted for just .3% and .1% of variance, respectively.

If, however, we consider the general skill effect as a baseline of performance, sequential multiple regression analyses can be performed to determine if the skill and context effects add anything to prediction of CRM OB ratings. Doing so, the skill effect accounts for just .1% more variability than the general skill effect alone, while the context effect accounted for 9.6% more variability than the general skill alone. Taken together, these results provide support for events sets, but very little for CRM skills, as individual units of LOE analysis.

Table 2. Squared **Zero-order** and Semi-partial Correlations Between Predictors and CRM OBs.

	Controlling for:					
Predictor	S	C	G	S & C	S & G	C & G
Skill (S)	**.378**	.017	.001	-	-	.003
Context (C)	.156	**.517**	.096	-	.098	-
General Skill (G)	.059	.015	**.436**	.001	-	-

Note. All values listed in table, except for values in bold, represent a semi-partial correlation between the CRM OB grades received and the constructed predictor variable in which the variable(s) listed along the top row of the table has been partialled out. If only one variable is listed, than it was the only one controlled for. Bold values represent the zero-order correlation of the constructed predictor variable with the CRM OB grades. Squared multiple R = .535.

Similar analyses were conducted in which technical OB ratings were predicted from the same three measures of general skill, skill, and context. These results are shown in Table 3. Overall, the three predictors accounted for 50.2% of the variability in technical OB ratings. Again, all three of the effects contributed significantly to prediction with the context effect emerging as the strongest predictor, accounting for 7.0% unique variance, and the general skill effect emerging as the weakest predictor in the model, accounting for less than .1%. The skill effect, however, accounted for 2.1% unique variance, somewhat more than in the CRM analysis.

Considering the general effect as the baseline of performance, sequential multiple regression analyses were also performed on the technical OB data. The skill effect accounted for 1.9% more variability than the general skill effect alone, while the context effect accounted for 6.7% more variability than the general effect alone. These results provide slightly more evidence for the validity of technical skills than that for CRM skills.

Table 3. *Squared **Zero-order** and Semi-partial Correlations Between Predictors and Technical OBs.*

	Controlling for:					
Predictor	S	C	G	S & C	S & G	C & G
Skill (S)	**.404**	.044	.019	-	-	.021
Context (C)	.098	**.458**	.068	-	.070	-
General Skill (G)	.028	.023	**.413**	.000	-	-

Note. All values listed in table, except for values in bold, represent a semi-partial correlation between the technical OB grades received and the constructed predictor variable in which the variable(s) listed along the top row of the table has been partialled out. If only one variable is listed, than it was the only one controlled for. Bold values represent the zero-order correlation of the constructed predictor variable with the technical OB grades. Squared multiple R = .502.

Several other analyses confirmed that technical OBs have relatively more skill structure than CRM OBs. Separate multiple regression analyses were conducted in which either CRM or technical OB ratings were predicted from the same Context and General Skill predictor variables as in the previous analyses along with several "skill" predictors. For prediction of CRM OB ratings, for example, scores for the Context and General Skill predictors were computed as before. In addition, a separate average was computed for a given crew's ratings on all OBs within each of the five CRM categories. Thus, CRM OB ratings were predicted from each of the five CRM categories, Context, and General Skill scores. Likewise, technical OB ratings were predicted from each of the four technical categories, Context and General Skill scores. It should also be noted that separate regression analyses were conducted on OBs within each CRM and technical skill. Thus, a total of nine regression analyses were run; one in which only CRM-1 OB ratings were predicted, one in which only CRM-2 OB ratings were predicted, and so on.

For the technical OBs, the same skill score turned out to be the best predictor other than Context. That is, the tech-1 average predicted tech-1 OB ratings better than tech-2, tech-3, or tech-4 averages. This was not true of the CRM OBs (see Table 4).

Table 4. *Squared Semi-partial Correlations Between Predictors (Specific Skills, Context (C), and General Skill (G)) of CRM and Technical OBs.*

	Predictor Variables						
Outcome Variable	C-1	C-2	C-3	C-4	C-5	C	G
CRM OB Ratings							
CRM-1	**.010**	.002	.000	.001	.001	.055	.001
CRM-2	.003	**.001**	.000	.002	.000	.097	.000
CRM-3	.000	.001	**.001**	.002	.005	.116	.002
CRM-4	.002	.003	.003	**.002**	.000	.104	.002
CRM-5	.001	.000	.004	.000	**.001**	.104	.002
	T-1	T-2	T-3	T-4		C	G
Technical OB Ratings							
Tech-1	**.044**	.001	.000	.000		.053	.001
Tech-2	.000	**.033**	.000	.000		.046	.000
Tech-3	.000	.000	**.002**	.000		.091	.000
Tech-4	.001	.000	.001	**.014**		.125	.001

Note. All values listed in table represent a semi-partial correlation between the CRM (designated at CRM-1 or C-1 through CRM-5 or C-5) or technical (designated at Tech-1 or T-1 through Tech-4 or T-4) OB grade received and the constructed predictor variable in which all other variables listed along the top row of the table have been partialled out.

Finally, a cluster analysis was performed on the technical OBs and the CRM OBs to determine if an underlying skill structure existed for either type of performance. Four groups emerged for the technical OBs consisting of the four skill categories. However, the cluster analysis on the CRM OBs failed to show similar rankings by categories. The CRM grades were most notably clustered around event sets, giving further proof of a context effect. The distribution of the technical grades also demonstrated this effect along with the category clustering.

Discussion

The basic finding from our study is that pilots' performance within CRM task categories lacks consistency. We found higher consistency of performance within flight contexts (i.e., phases of flight) than within CRM categories. These results question whether performance within standard CRM task categories (e.g., decision making, situation awareness) should be viewed as skills in the traditional sense of that term. These findings have implications for how CRM should be trained and assessed. However, before we discuss these issues, we wish to bring up some possible criticisms and cautions of the current findings.

First, in the present analyses a CRM category is operationally defined by the specific OBs used to assess performance within that category. Were the particular OBs used in the current study good measures of the CRM task performance they purported to measure? In defense of the OBs we can say that they were written by experienced evaluators from a major carrier with a history of assessing and training CRM. Hence, they are likely to be as good as any in the industry. Further, previous research has shown that behavioral markers are capable of assessing the skills and knowledge typically associated with CRM categories.

Second, perhaps the IEs were poor at discriminating among levels of performance associated with CRM tasks. Others have questioned whether IEs can grade CRM performance with the same accuracy as they do technical skills. Objective qualification standards (e.g., +- 10 deg heading difference) govern the grading of technical tasks, but are often lacking for CRM tasks. Admittedly, CRM by its very nature is more subjective. However, we have found that IEs' inter-rater reliability for grading CRM performance was as high as for grading technical performance in previous studies of training and calibration sessions (Goldsmith & Johnson, 2002). We applaud efforts to improve the reliability and validity of measures of CRM performance, but it is likely that the evaluations of human performance in the aviation industry is as good or better than any industry.

A third caveat is the low variability in the performance data. A high proportion (77%) of the grades were 3's, 4's and 5's. The effect sizes for correlation and regression analyses are mitigated by skewed distributions and low variance. Could the small CRM skill effects be due to restrictions on the distribution of grades? Perhaps, but arguing against this is the fact that we found with the same performance data context effects that were substantially higher than CRM skill effects.

Assuming the results from our study are valid, what can we claim about the psychological status of CRM and what are the implications for assessing and training it? First, it may be that CRM performance is a skill but that the traditional CRM categories used to evaluate it are incorrect. The division of CRM into decision making, planning, workload management, etc. may not reflect the categories that best differentiate true cognitive performance. One way of determining psychologically valid categories would be through cluster or factor analysis on large sets of performance data. What skill categories emerge from the empirical data? The cluster analysis we performed on the performance data in the present study resulted in a single CRM category.

A second possibility is that CRM is a skill but rather than composed of a set of subskill categories (e.g., decision making, situation awareness, etc.) it is best viewed as a unitary, general skill. This idea is supported by the results of the cluster analysis in the present study, and also by the fact that the category skill effect accounted for only 1.9% more variance in the grades than the general skill effect alone. These results suggest that the division of CRM into categories has little explanatory power. The implication is that pilots do vary on CRM performance, but rating them along distinct CRM subcategories does not help much in differentiating their performance. If true, then what particular CRM tasks are trained and assessed is of less importance than their receiving some CRM training.

Finally, CRM may not be a skill at all. Psychologists have long debated whether traits or situations best characterize human personality and performance (Epstein & O'Brien, 1985; Michel & Peake, 1982). Many have questioned the idea that people have enduring characteristics that manifest across the varied contexts of life. Rather our behavior is more a function of the particular situation we find ourselves in. This same idea seems to best explain the data in the current study on CRM performance. If true, then our training and assessing of CRM performance should focus more on sampling flight contexts than on CRM tasks.

References

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the mutitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Epstein, S., & O'Brien, E. J. (1985). The person-situation debate in historical and current perspective. *Psychological Bulletin, 98*, 513-537.

Goldsmith, T. E. & Johnson, P. J. (2002). Assessing and improving evaluation of aircrew performance, *International Journal of Aviation Psychology, 12*, 223-240.

Michel, W. & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89*, 730-755.

Salas, E., Wilson, K. A., Burke, C. S., & Wightman, D. C. (in press). Does CRM training work? An update, extension, and some critical needs. *Human Factors.*

Wiener, E. L., Kanki. B. G., & Helmreich, R. L. (Eds.). (1993). *Cockpit resource management*. San Diego, CA: Academic.

Welford, A. T. (1976). *Skilled performance: Perceptual and motor skills.* Oxford, England: Scott & Foresman.