Cite as: O'Connor, P., Campbell, J., Newon, J., Melton, J., Salas, E., & Wilson, K. (2008). Crew resource management training effectiveness: A meta-analysis and some critical needs. *International Journal of Aviation Psychology, 18(4),* 353-368.

Crew Resource Management Training Effectiveness: A meta-analysis and some critical needs.

Paul O'Connor[1], Justin Campbell[2], Jennifer Newon[3], John Melton[4], Eduardo Salas, [5]& Katherine A. Wilson[5, 6]

[1] School of Aviation Safety, Navy Aviation Schools Command
Pensacola, Florida USA

[2] University of West Florida, Pensacola, Florida, USA

[3]Loyola University Chicago
Chicago, Illinois USA

[4] Naval Service Training Command, Great Lakes, Illinois USA

[5]Department of Psychology
and
Institute for Simulation & Training
University of Central Florida
Orlando, Florida USA

[6] Now at William Lehman Injury Research Institute, University of Miami, Miami, Florida

Keywords: Crew Resource Management, team training, meta-analysis, training evaluation.

ABSTRACT

Empirical studies of Crew Resource Management (CRM) training effectiveness were subjected to meta-analysis. Sixteen CRM evaluation studies were found to fulfill the *a priori* criteria for inclusion in the meta-analysis. The metrics of CRM training effectiveness analyzed were: reactions, attitudes, knowledge, and behaviors. CRM trained participants responded positively to CRM (a mean of four on a five point Likert scale) and the training had large effects on the participants' attitudes and behaviors, and a medium effect on their knowledge. The findings from the meta-analysis are encouraging for the effectiveness of CRM training. However, there is a need for researchers, and reviewers, to be more rigorous about the data included in papers reporting CRM evaluation to allow effect sizes to be calculated.

INTRODUCTION

The aviation industry has been instrumental in developing training programs aimed at reducing human error and increasing the effectiveness of flight crews, a training commonly referred to as Crew Resource Management (CRM; Weiner, Kanki & Helmreich, 1993). CRM training can be defined as "*a set of instructional strategies designed to improve teamwork in the cockpit by applying well-tested tools (e.g., performance measures, exercises, feedback mechanisms) and appropriate training methods (e.g., simulators, lectures, videos) targeted at specific content (i.e., teamwork knowledge, skills, and attitudes)*" (Salas, Prince, et al., 1999: 163). Since its inception over 20 years ago, CRM training is now recommended by the major civil aviation regulators (e.g., Federal Aviation Authority, FAA; and Joint Aviation Authorities, JAA) and used by virtually all the large national and international airlines.

An introductory CRM course is generally conducted in a classroom for two or three days. Teaching methods include lectures, practical exercises, role playing, case studies, and video films of accident reenactments. To date, there is no standard set of competencies (i.e., knowledge, skills, and attitudes) that should be included as a part of CRM training, and wide array of concepts have been deemed "CRM" (Salas, Wilson, Burke, Wightman & Howse, 2006a). The literature tells us that CRM courses typically cover core topics such as teamwork, leadership, situation awareness, decision making, communication, and personal limitations (Flin & Martin, 2001; Salas et al., 2006a). Refresher training is typically a half or whole day course focusing on a specific CRM topic and is also required by the major civil aviation regulators (e.g. FAA, 2004; JAA, 2006). For flight deck crews, CRM skills then can be practiced and assessed in flight

simulator sessions known as line-oriented flight training (LOFT) and line operational evaluation (LOE).

CRM is one of the most widely applied techniques for providing team training to operations personnel in aviation, and consequently it has attracted the attention of other high-risk industries. Unsurprisingly, those that adopted it first were involved in the aviation industry: aviation maintenance, cabin crew, and air traffic control. However, CRM training is now beginning to extend beyond aviation into other high-risk industries (e.g. merchant navy, anesthesiology, emergency medicine, nuclear operator teams, the fire service, and oil production; see Flin, O'Connor & Mearns, 2002 for a review). Therefore, there is a need to evaluate the effectiveness of CRM training as it is applied in non-aviation domains. There is also a requirement to evaluate the effectiveness of CRM training in aviation to ensure it continues to have a positive effect on aircrew performance despite changes in aircraft design, operational conditions, emerging risks and aviator demographics.

CRM evaluation techniques

There is no simple answer to the question of whether CRM training can fulfill its purposes of increasing safety and efficiency (Helmreich, Merritt, & Wilhelm, 1999). The FAA (2004) states that for CRM training "*it is vital that each training program be assessed to determine if CRM training is achieving its goals. Each organization should have a systematic assessment process. Assessment should track the effects of the training program so that critical topics for recurrent training may be identified and continuous improvements may be made in all other respects*" (12: FAA, 2004).

Salas and Cannon-Bowers (1997) outline a number of principles for evaluating team training which have emerged over time. The recommended approach is one that is multi-faceted and considers several separate methods of assessment. The evaluation methods can be categorized into what is described by training researchers (e.g. Hamblin, 1974; Kirkpatrick, 1976) as different levels of training effects, ranging from individual to organizational indicators. Despite previous criticism (e.g. Alliger & Katzman, 1997), Kirkpatrick's (1976) hierarchy is still the most popular framework for guiding training evaluation (Salas & Cannon-Bowers, 2001; Salas, Wilson, Burke, Wightman, 2006). It provides a useful framework to assess the effects of a training intervention on an organization by considering training evaluations at multiple levels. Kirkpatrick's (1976) hierarchy consists of four different levels of evaluation: reactions, learning, behavior, and organizational impact (or results). These levels will be described, and the deductions of three literature reviews (O'Connor, Flin & Fletcher, 2002; Salas, Burke, Bowers & Wilson, 2001; Salas, Wilson, Burke, & Wightman, 2006) concerned with the effectiveness of CRM training will be summarized.

Level 1: Reactions. Evaluating reactions are the equivalent to measuring customer satisfaction. For example, did the participants like the training?, did the participants think the training was worthwhile? It is the most common type of evaluation data collected and is generally gathered using a paper-based questionnaire (O'Connor et al, 2002). The previous literature reviews have concluded that CRM participants generally have favorable reactions to the training (O'Connor et al, 2002; Salas et al, 2001; Salas, Wilson, Burke, Wightman, 2006).

Level 2: Learning. Learning is the second level in the hierarchy, and refers to "*the principles, facts, and skills which were understood and absorbed by the participants*" (Kirkpatrick, 1976: 11). Learning is made up of two components: attitudinal change and knowledge gains. The majority of data collected on participants' attitudes towards the topics covered in CRM training utilized a self-report survey. The most frequently used of these questionnaires is the cockpit management attitudes questionnaire (CMAQ; Helmreich, 1984). The CMAQ has been adapted for use in other high reliability industries in which CRM training is being applied (e.g. operating room management attitudes questionnaire). Previous reviews of CRM training evaluation studies have concluded that CRM training generally results in a positive change in CRM attitudes (O'Connor et al, 2002; Salas et al, 2001 Salas, Wilson, Burke, Wightman, 2006). Studies reporting a knowledge assessment after CRM training are reported less frequently than those in which attitude changes are evaluated. The most commonly reported assessment techniques were multiple choice tests (O'Connor et al, 2002). Overall, the literature reviews concluded there were increases in CRM knowledge of participants as a result of attending CRM training (O'Connor et al, 2002; Salas et al, 2001).

Level 3: Behavior. Evaluation of behavioral changes is the assessment of whether knowledge learned in training actually transfers to behaviors on the job, or a similar simulated environment. A widely used technique for assessing CRM skills in flight crew is to use an observational rating system known as behavioral markers to assess team behavior. Behavioral markers are "*a prescribed set of behaviors indicative of some aspect of performance*" (Flin & Martin, 2001: 96). They describe specific observable behaviors, not attitudes or personality traits (Klampfer, et al, 2001). A number of

different systems exist in the aviation industry (e.g. Targeted acceptable responses to generated events or tasks, TARGETS; Line/LOS checklist; Non-technical skills, NOTECHS). The conclusions drawn from the reviews of CRM evaluation studies in which behavioral measures were used are not in complete agreement. O'Connor et al (2002) and Salas et al (2001) concluded that there is evidence of changes in the behaviors of CRM trainees. However, Salas, Wilson, Burke, Wightman (2006) state that unlike the consistent positive results from those studies that examined reactions or learning, those studies that examined behavioral change were less conclusive.

Level 4: Organizational impact. This is the highest level of evaluation in Kirkpatrick's (1976) hierarchy. The ultimate aim of any training program is to produce tangible evidence at an organizational level, such as an improvement in safety and productivity. There are few studies in the literature reporting evaluations carried out at this level (O'Connor et al, 2002; Salas et al, 2001). Due to the low accident rates in the industries using CRM training, it is difficult to draw firm conclusions regarding the effect of CRM training on the organization.

Objectives of the Current Study

The purpose of this paper is to use meta-analyses techniques to evaluate the effectiveness of CRM training. The use of meta-analysis will improve upon previous narrative literature reviews through systematic application of quantitative procedures (Henry, Crawford, & Philips, 2005; Mullen & Rosenthal, 1985). With the proliferation of CRM literature, an empirical summation of the literature evaluating CRM effectiveness appears warranted.

METHOD

Literature search

Three previous narrative literature reviews of CRM training were the starting point for the review, the 58 studies reviewed by Salas et al (2001), the 48 studies reviewed by O'Connor et al (2002), and the 28 studies reviewed by Salas, Wilson, Burke, Wightman (2006). This process resulted in the identification of 74 CRM evaluation studies. Second, a computerized search of the literature from 1980 until 2006 was conducted utilizing PsycINFO, Google Scholar, Medline, and Defense Technical Information Center. Keywords for the computerized search of the literature were: *crew resource management*, *aircrew coordination training*, and *team co-ordination training*. No additional studies were identified in the computerized search of the literature. The 74 CRM evaluation papers detailed studies from industries including commercial aviation, military aviation, medicine, aviation maintenance, air traffic control, general aviation pilots, nuclear power generation, offshore oil production, and commercial shipping. A total of 32 papers were published in peer review journals, five in book chapters, 33 as conference papers, and four as technical reports.

Inclusion criteria

Three psychologists familiar with both training evaluation methods and meta-analysis made the decision, by consensus, as to whether a study met the *a priori* criteria for inclusion in the meta-analysis. For a study to be included in the meta-analysis, an evaluation had to be reported from at least one of the first three levels of Kirkpatrick's

(1976) evaluation hierarchy: reactions, learning (attitudes and knowledge), or behaviors. Also, if the data had been published in more than one study, then only one of the publications was included in the meta-analysis. No assessment was made at the organizational level due to the small number of studies, and diverse methods of evaluation methods at this level. The inclusion criteria for each of the three levels of Kirkpatrick's (1976) evaluation hierarchy are described below.

Inclusion criteria for reaction data.

Studies were included at this evaluation level if reactions were measured as a mean on a Likert scale, or as a percentage of participants who agreed with a certain statement. A measure was coded as a reaction when it pertained to any one of several perceptions of CRM training such as how much participants liked the training, what the felt about the training, how relevant they felt it was to their job, and how much of the information they felt they would use. If a mean score was given, it was necessary for a standard deviation to also be presented. If either a mean or percent agreement was given, the study also had to report the sample size. Also, the measurement had to be taken after the training.

Inclusion criteria for learning data.

For an assessment to be carried out of attitude change or knowledge gains, the study had to be either:

- based upon a comparison of a measurement taken prior to the CRM training and compared to an evaluation completed after the training course (a repeated measures design), or

- a comparison with a control group who had not received CRM training (a between groups design).

Studies comparing groups that had received different types of CRM training were excluded because they do not include a valid control group for the purposes of comparison. Pearson's *r* was used as the common effect size metric. Therefore, the studies had to contain sufficient information to allow Pearson's *r* to be calculated.

For an evaluation to be coded as an attitude measure it had to include some type of measure of attitudes that are conceptually or empirically related to CRM training using a tool such as the CMAQ. An evaluation was coded as a knowledge assessment if some type of knowledge measurement was completed (e.g. a multiple choice test).

Inclusion criteria for behavioral data.

For a study to be coded as an evaluation of behavior, there had to be an assessment carried out using some type of behavioral observation and coding system (e.g. TARGETS, Line/LOS checklist, NOTECHS). Further, as with the learning level, a comparison of performance had to be made before and after training, or with a control group who had not received any training, and contain sufficient information to allow Pearson's *r* to be calculated.

Although 74 CRM evaluation papers were identified, 58 papers did not meet the criteria for inclusion in the meta-analysis. Of the omitted studies, 40 reported insufficient statistical data for calculating study effects. Twelve studies did not make a comparison with participants who were naïve to CRM training. The remaining six studies included data already used in another study. Of the 16 studies included in the meta-analysis, 13

(81%) came from journals and 3 (19%) from conferences. A total of six (38%) with military aviators, four (25%) with medical personnel, three (19%) studies were carried out with civilian aviators, two with civil aviation students (13%) and one (6%) with offshore oil production personnel.

Calculating the effect size and analysis

A two-way random effects model, based upon methods developed by Hedges and Olkin (1985), was applied. Random effects models typically provide less biased indices of between studies variation compared to fixed-effects models which tend to overestimate the magnitude of mean effect sizes (National Research Council, 1992). The crucial statistical differences between the random and fixed effects models are that the random effects model typically leads to larger standard errors and confidence intervals (Henry et al, 2005). The random effects model assumes a normal distribution of effect sizes, with variations in effect sizes attributable to sampling error.

Studies that utilized between groups and repeated measures designs were analyzed independently. This separation was necessary to address the substantive difference between the two research designs (Ray & Shadish, 1996; Morris & DeShon, 2002). Five of the studies included in the meta-analysis utilized both between and repeated measures designs, which made it permissible for each study to contribute separate, independent effects to the calculation of the between and within groups mean effects.

With respect to the standardization of effect sizes, it would have been desirable to simply pool means and standard deviations to create meta-analysis of Cohen's *d* to represent the average between or repeated measures differences in CRM attitudes,

knowledge, and behaviors. However, means and standard deviations were not reported in several instances, in some cases leaving only p-values or effect sizes. Therefore, meta-analysis techniques for correlation coefficients, which are more amenable to a wide array of effects information (Rosenberg, Adams, & Gurevitch, 2000), were used to calculate the mean-effect using the MetaWin (version 2) software (Rosenberg et al., 2000). The mean effects were transformed to, and reported, as Cohen's *d* to remain consistent with the majority of effects that represented mean comparisons.

Cohen's *d* describes the standardized difference between two group means. Effect direction was reversed when necessary to ensure that positive effects (both *r* and *d*) indicate an improvement in scores on CRM attitudes, knowledge, or behavior measures. Confidence intervals (CIs, 95%) were also calculated. CIs are used to assess the accuracy of the estimate of the mean effect size (Whitener, 1990). A CI that does not bound zero is indicative of a significant effect the $p < .05$ level.

The *Q*-test (Hedges & Olkin, 1985) was used to estimate the degree of heterogeneity of the effects. A significant *Q* is associated with differences between the effects contributing to the mean. In contrast, a non-significant *Q* is indicates a lack of substantial differences between the effects contributing to the mean, once sampling error has been removed.

RESULTS

Reactions

Seven studies included data that was suitable to code for reaction measures. All of the studies included used a post-training paper and pencil survey. The data were reported in the form of a five point Likert Scale (see Table 1). Participants gave a high mean

weighted reaction to CRM training. On average participants gave the usefulness of CRM training a four out of five possible points (mean= 4.18). Therefore, participants in CRM training feel that the training is useful.

Table 1. Mean score (from 1 not useful to 5 very useful) and standard error for post-CRM training reactions.

Author	Population	Mean	Standard error
Baker et al (1993)	41 U.S. Navy helicopter pilots	4.37	0.12
Baker et al (1991)	36 U.S. military instructor pilots	4.28	0.13
Baker et al (1991)	46 U.S. military student pilots	4.21	0.12
Holzman et al (1995)	31 anesthesiologists (attending)	3.88	0.20
Holzman et al (1995)	37 anesthesiologists (house officers)	4.24	0.14
Kurrek et al (1996)*	59 anesthesiologistS	4.00	0.13
O'Connor & Flin (2003)	77 North Sea offshore oil workers	3.85	0.05
Salas, Fowlkes et al (1999)	67 U.S. Navy aircrew	4.13	0.18
Stout et al (1997)	20 U.S. Navy pilots	4.66	0.12
Combined		**4.18**	

* The scale of this study was reversed to be consistent with the other studies analyzed.

Attitudes

The most commonly used measures of attitude were either the Cockpit Management Attitude Questionnaire (CMAQ) or Flight Management Attitude Questionnaire (FMAQ). A meta-analysis was carried out using the three studies (see the top of Table 2) that compared the attitudes of CRM trained personnel versus an un-trained control group. A mean effect size of $d_{mean} = .94$ resulted, with a confidence interval that does not overlap zero (see the bottom of Table 2). Cohen (1988) provides guidelines for interpreting *d* effect sizes: 0.2 or less is a small effect, 0.5 is a medium effect, and 0.8 or greater is considered a large effect. These findings indicate a large and significant improvement in

the attitudes of training participants to the topics covered in CRM when compared to those who did not receive CRM training.

Table 2. Study descriptors, effect information and meta-analysis of between group effects.

Study	**Sample and Ntotal** [a]	**r**
Attitudes		
Fisher, Phillips & Mather (2000).[b]	66 US air medical crew.	0.26
Morey, Simon, Jay, Wears, et al. (2002).	1,058 hospital emergency departments staff	0.49
Stout, Salas, & Fowlkes (1997).	42 U.S. Navy pilots	0.46
Knowledge		
O'Connor & Flin (2003).	86 North Sea offshore oil production platform personnel	0.04
Salas Fowlkes, et al (1999).	67 U.S. Navy aircrew	0.54
Stout, Salas, & Kraiger (1996).	12 U.S. Navy helicopter pilots	0.04
Stout, Salas, & Fowlkes (1997).	20 U.S. Navy pilots	0.40
Behavior		
Clothier (1991).	Major U.S airline (2,110 crew members)	0.82
Connolly & Blackwell (1987).	29 aeronautical science student	0.83
Morey, Simon, Jay, Wears, et al. (2002).	1,058 hospital emergency departments staff	0.81
Shapiro, Morey, Small, Langford, et al (2004).	16 medical teams	0.18
Stout, Salas, & Fowlkes (1997).	32 U.S. Navy pilots	0.52

Mean Effects				**95% CI of mean**		
	k	**n**	**d_{mean}**	**Lower**	**Upper**	**Q**
Attitude	1,166	3	0.94	.18	1.86	1.78, n.s.
Knowledge	185	4	0.59	-.41	1.74	2.51, n.s.
Behavior	3,245	5	1.18	-.25	3.18	2.74, n.s.

Notes. a: N represents the total of CRM trained plus CRM non-trained. b: The scale of this study was reversed to be consistent with the other studies analyzed.

A meta-analysis of the six studies that assessed participants' attitudes to CRM training using a repeated measures design also resulted in a large, positive effect size (d_{mean} = .85; see the bottom of Table 3). However, substantial variation in effect sizes across these studies (.13 to .71; see Table 3) contributed to a large CI which bounds zero.

Therefore, the large mean difference must be interpreted with caution. Despite the large variation in the repeated measures effects for attitudes, the Q-value was not significant, possibly due to the relatively small number of effects that contributed to this analysis. The small number of between-group effects would also have impacted the Q test for those effects, which was also not significant. Collectively, the magnitude of the between-groups and repeated-measures mean effects indicate improvement in CRM attitudes after training, albeit the evidence relies on just a few effect sizes that are highly variable.

Table 3. CRM descriptors, effect information and meta-analysis of utilizing a repeated measures design.

Author(s)	**Sample**	**r**
Attitudes		
Gregorich, Helmreich & Wilhelm (1990).	673 flight crews	0.14
Leedom & Simon (1995).	32 U.S army helicopter aviators	0.53
O'Connor & Flin (2003).	77 North Sea offshore oil production platform personnel	0.13
Morey, Simon, Jay, Wears, et al. (2002).	684 hospital emergency departments staff	0.71
Salas Fowlkes, et al (1999).	69 U.S. Navy aircrew	0.32
Stout, Salas, & Fowlkes (1997).	40 U.S. Navy pilots	0.37
Behavior		
Connolly & Blackwell (1987).	16 aeronautical science student	0.83
Lassiter, Vaughn, Smaltz, Morgan, & Salas (1990).	90 undergraduate students	0.14
Leedom & Simon (1995).	31 U.S army helicopter aviators	0.62
Morey, Simon, Jay, Wears, et al. (2002).	684 hospital emergency departments staff	0.73

Mean Effects				**95% CI of mean**		
	k	**n**	**d_{mean}**	**Lower**	**Upper**	**Q**
Attitude	1,575	6	0.85	-.14	2.03	2.01, n.s.
Behavior	821	4	1.60	-.09	4.14	2.52, n.s.

Knowledge

Only between-groups comparisons were made with respect to mean differences in CRM knowledge pre and post CRM training. The four between group knowledge effects produced a large mean effect size (see Table 2), supporting an improvement in CRM knowledge for CRM training participants. However, again, the 95% CI bounds zero, therefore this effect was not significant.

Behavior

According to Cohen's (1988) guidance, effect sizes at the behavioral level were large, a finding that applied to both the five between groups effects and the four repeated measures effects (see Tables 2 and 3). This finding suggests substantial improvement in CRM behaviors associated with undergoing CRM training. Nevertheless, both types of effects produced 95% CI that bounded zero, again making it difficult to eliminate effect sampling error as a possible explanation for the finding.

DISCUSSION

General findings

The reactions to CRM training were positive, large effects of CRM training were found for attitudes and behaviors, and a medium effect size was found for knowledge—congruent with other reviews (e.g., Salas et al., 2001; Salas, Wilson, Burke, Wightman, 2006; O'Connor et al, 2002). The magnitude of the effect sizes was greater than other meta-analyses of training interventions. Arthur, Bennett, Edens, and Bell (2003) conducted a meta-analysis of individual training interventions to examine the effect of

training on the reactions, learning, and behaviors of trainees (team training interventions, such as CRM training were explicitly not included in this meta-analysis). For those training techniques that were designed to affect cognitive processes and interpersonal relationships (as is the aim of CRM training, Gregorich & Wilhelm, 1993), Arthur et al (2003) found an effect size of d=0.26 for learning, and d= 0.30 for behaviors. Similarly, Guzzo, Jette, and Katzell (1985) reported a mean effect size of 0.44 for all psychologically based interventions.

In the meta-analysis reported in this paper, except for the comparison of attitudes for the studies that utilized a between groups design, the 95% CI of the effect sizes bounded zero. Therefore, it is not possible to conclude that the effect sizes were significant. Nevertheless, despite the lack of significant effects (Arthur et al, 2003 also reported confidence intervals for training effects that bounded zero), the findings are worth commenting on in more detail. Each of the levels of evaluation are discussed below.

Reactions

Participants had an overall positive reaction to CRM training, with a mean of four on a five point Likert scale. The finding that participants had positive reactions to CRM training is consistent with the conclusions from the three CRM training evaluation literature reviews (O'Connor, et al 2002; Salas, et al, 2001; Salas, Wilson, Burke, Wightman, 2006).

Knowledge

A medium effect of CRM training was found for knowledge evaluation. Of the four studies included in the meta-analysis in which a knowledge assessment was carried out, three used multiple choice assessment, and one used a case study to ascertain whether the participants had improved in their ability to identify the human factors causes of the incident as a result of the training (O'Connor & Flin, 2003). The effect of CRM training on knowledge has not been widely reported in the literature. For example, only 15% of the CRM evaluation studies identified by O'Connor et al (2002) reported an assessment of the knowledge of participants.

Attitudes

A large effect size of CRM training was found for attitudes. Only one of the studies (Morey et al, 2002) included in the meta-analysis did not use a questionnaire that was based upon the CMAQ to make an assessment of attitude change as a result of CRM training. The CMAQ is a well established training, evaluation and research tool developed to assess the effects of CRM training for flight crews. However, depending on the content of the CRM training, it is possible that the CMAQ will not measure the range of attitudes that should have changed. The CMAQ was developed solely to assess pilots' attitudes regarding *interpersonal components* of the flight crew's job performance (Gregorich, Helmreich, & Wilhelm, 1990; Helmreich, 1984). It does not address attitudes to the *cognitive* aspects of the role of the flight crew such as situation awareness, decision making, and workload management. The cognitive aspects of CRM are crucial components to the training, and recommended for inclusion by the major commercial

aviation regulatory bodies (FAA, 2004; JAA, 2006). Therefore, arguably, the CMAQ is only evaluating a subset of the attitudes that are being addressed by CRM training.

Behavior

A large effect was found at the behavioral level, with a greater effect size than found for attitudes or knowledge. However, the conclusion of a significant effect of CRM training on behaviors cannot be made as the 95% CI bounded zero. Nevertheless, there are a number of possible explanations for the large effect size. Firstly, in those studies in which a measure of behavior was carried out, the training itself tended to have a strong focus on changing behaviors (e.g. Leedom & Simon, 1995; Morey et al, 2002; Stout, Salas & Fowlkes, 1997). Secondly, in some of the studies participants were given the opportunity to practice the behaviors they had learned in the classroom in simulators as part of the CRM training course (Leedom & Simon, 1995; Stout et al, 1997). A combination of lectures, the opportunity to practice desirable behaviors, and feedback regarding performance is a well established mechanism for delivering effective training (e.g. Baldwin & Ford, 1988, Bandura, 1977). Further, recommendations for using behavioral techniques to conduct CRM training have been made many times in the literature (e.g. Salas, Rhodenizer & Bowers, 2000; Salas, Wilson, Burke, Wightman, & Howse, 2006b). Third, there may be some experimental design considerations that could have inflated the effect size. It is a possibility that the participants under observation may have temporarily changed their behavior as they knew they were under observation (a Hawthorne effect). It was also not reported in any of the studies whether the observers had been blinded to whether the people they were observing were a control group, pre-, or post CRM training.

Methodological limitations

The use of meta-analysis leads to a gain in precision at the cost of selectivity (Guzzo, Jackson & Katzell, 1987). A total of 55 CRM evaluation studies were not included in the meta-analysis due to insufficient data. This may lead to doubts about the findings of the meta-analysis. However, as the findings are in general agreement with the literature reviews carried out by O'Connor et al (2002), Salas et al (2001), and Salas, Wilson, Burke, Wightman (2006), the authors are confident that the studies included in the meta-analysis are representative of the larger body of CRM evaluation research.

Nevertheless, the low number of studies included in the meta-analysis represents a large weakness of the present effort. The vast majority of the studies examined could not be included in the meta-analysis due to insufficient data being reported to allow effect sizes to be calculated. Researchers must ensure that CRM evaluation studies provide detailed information on the training that was given, the evaluation metrics used, and the results of the evaluation. The American Psychological Society (APA) states that "*it is almost always necessary to include some index of effect size or strength of relationship*" (APA, 2001: 25). As CRM training becomes prevalent in other industries, it is crucial that CRM training effectiveness is assessed. However, until researchers and reviews become more rigorous in the reporting of data from CRM evaluation studies, it will not be possible to complete a more inclusive meta-analysis.

According to Flores and Crepaz (2004), standard reporting of an intervention's quality of methods should include: thoroughly describing intervention activities, using comparable evaluation measures at pre-intervention and post-intervention, reporting results in detail for each group, reporting rates of refusal to participate, including a

comparison group and demonstrating study group equivalence, clearly describing study group assignment, use of appropriate statistical controls, and collecting follow-up data from at least three months post-training.

CONCLUSION

Due to a lack of sufficient data, it was only possible to include a subset of the published CRM evaluation studies in the meta-analysis. As CRM training gains continued popularity in high risk communities beyond aviation, it is imperative that researchers and reviewers be more rigorous about reporting CRM training evaluations. The communities implementing CRM training must demand it. We need a mandate compelling researchers to conduct more robust research, access to the data (e.g., open cockpits, operating rooms) and resources (e.g., money) to make this a reality (Salas et al., 2006a). Although a significant effect of CRM training was only found at the attitude level, the findings from the meta-analysis of CRM effectiveness are encouraging. It is our hope that this paper will encourage continued research in this area, as well as highlight the importance of reporting statistical data so that we can truly examine CRM training's effectiveness.

Acknowledgements

The research and development reported here was partially supported by the FAA Office of the Chief Scientific and Technical Advisor for Human Factors, AAR-100. Dr. Eleana Edens was the contracting officer's technical representative at the FAA. All opinions stated in this paper are those of the authors and do not necessarily represent the opinion or position of the School of Aviation Safety, University of West Florida, Loyola

University, Naval Service Training Command, University of Central Florida, or the Federal Aviation Administration.

REFERENCES

Alliger, G. M. & Katzman, S. (1997). When training affects variability: beyond the assessment of mean differences in training evaluation. In J. K. Ford (Ed.), *Improving training effectiveness in work organizations* (pp. 223-246). Mahwah, NJ: Erlbaum.

American Psychological Association (2001). *Publication manual of the American Psychological Association*. Washington, DC: Author.

Arthur, W., Bennett, W., Edens, P.S., & Bell, S.T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88*, 234-245.

References marked with an asterisk indicates studies included in the meta-analysis.

*Baker, D., Bauman, M. & Zalesny, M. D. (1991). Development of aircrew coordination exercises to facilitate training transfer. In R. Jensen (Ed.) *Proceedings of the 6th International Symposium on Aviation Psychology* (pp. 314-319), Columbus, Ohio.

*Baker, D., Prince, C., Shrestha, L., Oser, R., & Salas, E. (1993). Aviation computer games for crew resource management training. *The International Journal of Aviation Psychology*, 3(2), 143-156.

Baldwin, T.T. & Ford, J.K. (1988). Transfer of training: a review and direction for future research, *Personnel Psychology, 42*, 331-342.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

*Clothier, C. (1991). Behavioral interactions across various aircraft types: Results of systematic observations of line operations and simulations. In R. Jensen (Ed.) *Proceedings of the 6th International Symposium on Aviation Psychology*, (pp. 332-337). Ohio State University, Columbus, OH.

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

*Connolly, T. J., & Blackwell, B. B. (1987). A simulator approach to training in aeronautical decision making. In R. Jensen (Ed.) *Proceedings of the 4th International Symposium on Aviation Psychology*, (pp. 251-258), Columbus, Ohio.

Federal Aviation Authority (2004). *Advisory Circular No 120-51E: Crew resource management training*. Washington, DC: Author.

*Fisher, J., Phillips, E., & Mather, J. (2000). Does crew resource management training work? *Air Medical Journal, 19,* 137-139.

Flin, R. & Martin, L. (2001). Behavioral markers for Crew Resource Management: A review of current practice. *International Journal of Aviation Psychology, 11,* 95-118.

Flin, R., O'Connor, P. & Mearns, K. (2002) Crew Resource Management: Improving Safety In High Reliability Industries. *Team Performance Management, 8,* 68-78.

Flores, S.A. & Crepaz, N. (2004). Quality of study methods in individual- and group-level HIV interventions research: critical reporting elements. *AIDS education and prevention, 16,* 341-352.

*Gregorich, S. E., Helmreich, R. L., & Wilhelm, J. A. (1990). The structure of cockpit management attitudes. *Journal of Applied Psychology, 75,* 682-690.

Gregorich, S. E. & Wilhelm, J. A. (1993). Crew resource management training assessment. In E. L. Wiener, B. G. Kanki & R. L. Helmreich (Eds.), *Cockpit Resource Management* (pp. 173-196). San Diego: Academic Press.

Guzzo, R.A., Jackson, S.E., & Katzell, R.A. (1987). Meta analysis. *Research in Organizational Behavior, 9,* 407-442.

Guzzo, R.A. Jette, R.D. & Katzell, R.A. (1985). The effects of psychologically based intervention programs on worker productivity. *Personnel Psychology, 38,* 275-291.

Hamblin, A. C. (1974). *Evaluation and control of training*. London: McGraw Hill.

Hedges, L.V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Helmreich, R. (1984). Cockpit management attitudes. *Human Factors, 26,* 583-589.

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of crew resource management training in commercial aviation. *International Journal of Aviation Psychology, 9,* 19-32.

Henry, J.D., Crawford, J.R., Philips, L.H. (2005). A meta-analytic review of verbal fluency deficits in Huntington's disease. *Neuropsychology, 19,* 243-252.

*Holzman, R. S., Cooper, J. B., Gaba, D. M., Philip, J. H., Small, S. D., & Feinstein, D. (1995). Anesthesia crisis resource management: Real-life simulation training in operating crises. *Journal of Clinical Anesthesia, 7,* 675-687.

Joint Aviation Authority (2006). *JAR-OPS, 1 Subpart N, Crew resource management flight crew (amendment 12).* Hoofddorp, Netherlands: Author.

Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig & L. R. Bittel (Eds.), *Training and development handbook* (pp. 18.1-18.27). New York: McGraw Hill.

Klampfer, B., Flin, R. Helmreich, R, Häusler, R., Sexton, B., Fletcher, G., et al. (2001) *Enhancing performance in high risk environments: Recommendations for the use*

of behavioural markers. Report from the behavioural markers workshop, Zürich, June. Berlin: Damler Benz Foundation.

*Kurrek, M. M. & Fish, K. J. (1996). Anesthesia crisis resource management training: An intimidating concept, a rewarding experience. *Canadian Journal of Anesthesia*, *43*, 430-434.

*Lassiter, D. L., Vaughn, J.S., Smaltz, V.E., Morgan, B.B., & Salas, E. (1990). A comparison of two types of training interventions on team communication performance. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 1372-1376). Santa Monica, CA: Human Factors and Ergonomics Society.

*Leedom, D. K., & Simon, R. (1995). Improving team coordination: A case for behavior -based training. *Military Psychology, 7*, 109-122.

*Morey, J. C., Simon, R., Jay, G. D., Wears, R. L., Salisbury, M., Dukes, K. A. et al. (2002). Error reduction and performance improvement in the emergency department through formal teamwork training: Evaluation results of the MedTeams project. *Health Services Research*, 37, 1553-1581.

Morris, S.B. & Deshon, R.P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105-125.

Mullen, B. & Rosenthal, R. (1985). *Advanced BASIC meta-analysis*. New York: Sage Publications.

National Research Council (1992). *Combining information: Statistical issues and opportunities for research*. Washington DC: National Academy Press.

O'Connor, P., & Flin, R. (2003). Crew resource management training for offshore oil production teams. *Safety Science, 41*, 591-609.

O'Connor, P., Flin, R. & Fletcher, G. (2002) Methods used to evaluate the effectiveness of CRM training: A literature review. *Journal of Human Factors and Aerospace Safety, 2*, 217-234.

Ray, J.W. & Shadish, W.R. (1996). How interchangeable are different estimators if effect size? *Journal of Consulting and Clinical Psychology, 64*, 1316-1325.

Rosenberg, M. S., Adams, D. C., & Gurevitch, J. (2000). *Meta Win 2.0*. Sunderland, MA: Sinauer.

Salas, E., Burke, C. S., Bowers, C. A., & Wilson, K. A. (2001). Team training in the skies: Does crew resource management (CRM) training work? *Human Factors, 41*, 161-172.

Salas, E. & Cannon-Bowers, J. A. (1997). Methods, tools and strategies for team training. In M. Quinones & E. Ehrestein (Eds.), *Training for a rapidly changing workplace: Applications in psychological research* (pp. 291-322). Washington DC: American Psychological Association Press.

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52*, 471-499.

*Salas, E., Fowlkes, J. E., Stout, R. J., Milanovich, D. M., & Prince, C. (1999). Does CRM training improve teamwork skills in the cockpit?: Two evaluation studies. *Human Factors, 41*, 326-343.

Salas, E., Prince, C., Bowers, C., Stout, R., Oser, R. L., & Cannon-Bowers, J. A. (1999). A methodology for enhancing crew resource management training. *Human Factors, 41*, 161-172.

Salas, E., Rhodenizer, L., & Bowers, C.A. (2000). The design and delivery of CRM training: Using all of the resource available. *Human Factors, 42*, 490-511.

Salas, E., Wilson, K.A, Burke, C.S., & Wightman, D.C. (2006). Does CRM training work? An update, extension and some critical needs. *Human Factors, 14*, 392-412.

Salas, E., Wilson, K. A., Burke, C. S., Wightman, D. C., & Howse, W.R. (2006a). Crew resource management training research, practice, and lessons learned. In R. C. Williges (Ed.), *Review of human factors and ergonomics* (Vol. 2, pp. 35-73). Santa Monica, CA: Human Factors and Ergonomics Society.

Salas, E., Wilson, K.A, Burke, C.S., Wightman, D.C., & Howse (2006b). A checklist for crew resource management training. *Ergonomics in Design*, *14*, 6-15.

*Shapiro, M.J., Morey, J.C., Small, S.D., Langford, V., Kaylor, C.J., Jagminas, L., et al. (2004). Simulation based teamwork training for emergency department staff: does it improve clinical team performance when added to an existing didactic teamwork curriculum? *Quality and Safety in Health Care, 13*, 417-21.

*Stout, R. J., Salas, E., & Fowlkes, J. E. (1997). Enhancing teamwork in complex environments through team training. *Group Dynamics: Theory, research, and practice, 1*, 169-182.

*Stout, R. J., Salas, E., & Kraiger, K. (1996). The role of trainee knowledge structures in aviation psychology. *The International Journal of Aviation Psychology, 7,* 235-250.

Whitener, E.M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75,* 315-321.

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