2003-01-3062

Airline Pilots' Experiences in and Reactions to their Check Rides: Results from a Nationwide, Representative Survey

David P. Baker, J. Matthew Beaubien, & Gonzalo Ferro
American Institutes for Research

ABSTRACT

While a substantial body of research has explored the effectiveness of airline pilot training programs, few studies have examined the check rides that occur at the end of training. To address this critical gap, we conducted a nationwide, representative survey of commercial airline pilots. In this paper, we explore their reactions to maneuver validations (MVs) and Line Operational Evaluations (LOEs). On average, the respondents rated both types of checking procedures favorably. Moreover, despite having a representative sample, reliable scales, and a high degree of statistical power, we found no practically or statistically significant differences between the perceived effectiveness of MVs and LOEs. The data suggest that airline pilots perceive both types of check rides as being equally effective. Implications and directions for future research are discussed.

INTRODUCTION

Despite sensationalistic news reports that might lead one to believe otherwise, commercial aviation remains the safest form of mass transportation in the United States. In fact, the probability of surviving any given flight is greater than 99.99% (National Transportation Safety Board, 1994). This impressive safety record is due in part to the rigorous training and evaluation procedures that commercial pilots undergo.

Much has been written on the effectiveness of airline pilot training programs. For example, Helmreich and colleagues demonstrated the need to supplement pilots' technical flying skills with Crew Resource Management (CRM) skills such as communication, decision-making, and situational monitoring (Helmreich & Foushee, 1993; Helmreich, Weiner, & Kanki, 1993). Some years later, Salas and colleagues' demonstrated that pilots who have received CRM training show more positive attitudes towards teamwork, increased knowledge of teamwork principles, and superior performance on simulated flights than pilots who have not received such training (Salas, Burke, Bowers, & Wilson, 2001). Finally, recent research by Baker and colleagues has shown that training programs which integrate CRM principles throughout the training curriculum are perceived as more useful than stand-alone CRM courses (Baker, Beaubien, & Mulqueen, 2002; Beaubien & Baker, 2002).

Unfortunately, there is a dearth of information concerning the types of checking or evaluation procedures that occur at the end of training. As a practical matter, evaluation is an inherent component of training. At the individual level, check rides – such as maneuver validations or Line Operation Evaluations – are used to certify the airworthiness of individual pilots. At the aggregate level, evaluation studies are used to assess the extent to which training goals are realistic, the training content and instructional techniques are sufficient to achieve the intended goals, and the training results in a quantifiable return on the organization's investment (Cannon-Bowers, Prince, Salas, Owens, Morgan, & Gonos, 1989).

The focus of this paper is on the individual-level, post-training evaluations which are also known as "check rides." There are two major types of check rides in commercial aviation. Under traditional pilot training procedures (14 CFR Part 121), the check ride is known as a maneuver validation (MV). The MV requires each pilot to individually perform a series of critical maneuvers that are graded according to FAA and airline standards. Although the MV is conducted in a full crew environment, each pilot must demonstrate his or her individual skill proficiency on every maneuver. Typically, there is little or no emphasis on CRM skills during a maneuver validation. Moreover, by focusing on individual maneuvers with few contextual cues, the MV may have high physical fidelity but low psychological fidelity (Boehm-Davis, Holt, & Hansberger, 1999).

The Advanced Qualification Program (AQP) is a voluntary alternative to traditional pilot training and checking procedures under 14 CFR Part 121 (Federal Aviation Administration, 1991, 1990b). AQP differs substantially from Part 121 training along a number of dimensions. For example, whereas the content of Part 121 training is standardized across carriers, AQP-participating carriers have substantial latitude to tailor the training content to their specific needs. Moreover, AQP requires that CRM principles be integrated throughout all phases of pilot crew training and evaluation. Finally, the AQP check ride, which is known as a Line Operational Evaluation (LOE), is designed to simulate typical flight operations from takeoff to landing.

During the LOE, the pilots are evaluated as a crew, and substantial emphasis is placed on assessing both their technical and CRM skills (Federal Aviation Administration, 1991, 1990a, 1990b). As a result, the psychological fidelity of the LOE should be substantially higher than that of a traditional maneuver validation.

The purpose of this study was to assess the relative effectiveness of different approaches to checking pilot performance at the end of training: the maneuver validation (MV) and the Line Operational Evaluation (LOE). Because the LOE provides greater contextual cues and integrates CRM skills with technical skills, it should simulate typical line operations more accurately than a traditional maneuver validation. Therefore, we hypothesized that pilots would rate the LOE as more useful than the MV. The results presented below are part of a much larger survey of airline pilots' experiences in and reactions to their professional training (Baker et al., 2002).

METHOD

PARTICIPANTS

The survey participants included 10,166 line-qualified pilots from 24 U.S. airlines. There were roughly equal numbers of Captains (49.1%) and First Officers (44.4%). The participants included both highly seasoned veterans and relative novices. A sizeable number reported that they had logged over 14,000 hours in commercial and military aircraft (25.8%). However, most reported having logged between 2,000 and 14,000 hours (72.9%). A handful reported having flown fewer than 2,000 hours (1.3%).

Prior to joining their current airline, many had previously flown for regional carriers (30.5%), supplemental or cargo carriers (13.2%), the military (52.2%), private companies or charter carriers (34.1%), or for other types of flight operations (33.4%). Fewer had flown for other major (11.3%) or national carriers (7.6%).

Check rides are not required at the end of every pilot crew training event (for example, annual recurrent training may be split into two or more separate "events"). Therefore, the study's primary hypothesis was tested using only those pilots who had received a check ride during their most recent training, and who provided valid data for each of the covariates (n=6,952 pilots or 68.4% of the respondents who provided usable data).

MATERIALS

The materials included a cover letter, a survey, and a follow-up postcard. The cover letter was printed on union letterhead, signed by the union president, and personally addressed to each pilot. The survey addressed a variety of issues, including pilots' general reactions toward their training. It also included specific questions on check rides and Crew Resource Management (CRM) training. All of the survey questions used a 5-point Likert scale with anchors that ranged from 1 (strongly disagree) to 5 (strongly agree).

MEASURES

Independent and Dependent Variables

The independent variable was the type of check ride (maneuver validation or Line Operational Evaluation) that the pilots received. The dependent variable was a 5-item measure that assessed the check ride's perceived usefulness. The items were as follows: "The check ride realistically represented line operations," "The check ride prepared me to fly the line," "The check ride provided useful feedback about my performance," "The check ride was administered in a fair and impartial manner," and "The check ride was a realistic evaluation of my piloting skills" (α=.85).

Covariates

To ensure that the survey results were not contaminated by extraneous factors, we included several covariates in the analysis. These include the pilots' rank (Captain vs. First Officer) and number of flight hours. We also included two additional scales from the survey. The first scale contained 4 items that assessed the pilots' personal experiences in CRM training. The scale items were as follows: "CRM training provided useful feedback about my performance," "CRM training covered important issues in current line operations," "The objectives of CRM training were clear," and "CRM training prepared me to fly the line" (α=.89). High scores indicated positive experiences in CRM training.

The second scale contained 4 items that assessed the pilots' recognition of the balance between training and evaluation. The scale items were as follows: "In general, the emphasis was on training as opposed to evaluation," "In general, the purpose of evaluations was clear," "In general, evaluations provided useful feedback about my performance," and "In general, it was clear when I was being evaluated versus being trained" (α=.82). High scores indicated a positive and realistic understanding of the tightly coupled relationship between training and checking.

Because the procedures for developing and administering check rides are relatively standardized (Federal Aviation Administration, 1991), we did not expect the perceived utility of check rides to vary as a function of the pilots' rank or number of flight hours. However, we did expect that favorable reactions to CRM training and a healthy understanding of the balance between training and evaluation would be positively correlated with pilots' reactions to their check rides.

PROCEDURE

Survey development involved several steps: document review, focus groups, item development, and pre-testing.

We began by reviewing the aviation psychology, training, and human factors research literatures to identify important issues in pilot crew training and debriefing. We then conducted a series of focus groups with pilots to better understand which issues personally affect them. Based on our findings, we developed a series of survey items. We then pre-tested the items with four samples of airline pilots. After each pre-test, we revised the instrument as necessary. Pre-testing continued until no more substantive changes were required to either the survey content or layout.

Membership lists from 3 major pilot unions served as our sampling frame. We used stratified random sampling to select one-half of the pilots (n=30,732) from 24 U.S. carriers that was representative in terms of airline, aircraft type, and pilot rank. The survey was administered via U.S. mail to the pilots' home addresses. To increase the response rate, we sent a follow-up postcard 2 weeks after the initial survey mailing (Yammarino, Skinner, & Childers, 1991).

RESULTS

RESPONDENT REPRESENTATIVENESS

Prior to conducting any statistical analyses, we performed a series of data screening and checking procedures to ensure the quality of the data (Tabachnick & Fidell, 1993). Because the survey was machine scored, we encountered relatively few problems. Next, we calculated the survey respondents' representativeness vis-à-vis the intended population. The differences between the sample and usable response proportions (per carrier) were generally less than 1%, and never exceeded 4%. These differences were extremely small, and obviated the need for weighting the survey results.

MEASUREMENT EQUIVALENCE

Because a maneuver validation (MV) is conceptually distinct from a Line-Oriented Evaluation (LOE), we wanted to ensure that both groups of pilots were responding to the survey items in roughly the same manner. We did this by conducting a multiple-groups confirmatory factor analysis (MGCFA) on the dependent variable. The first step was to establish a one-factor "baseline model" that included all of the respondents. To establish the scale for the latent variable, the factor loading for the first item was constrained to 1.0 (Byrne, 1998).

Over the years, several indices of model fit have been proposed (Bollen & Long, 1993). These include the Chi-Square (χ^2) index, the Tucker-Lewis Index (TLI), the Comparative Fit Index (CFI), and the Standardized Root Mean Square Residual (SRMR). In general, the χ^2 index provides a reasonable measure of model fit. However, it is usually significant (i.e., indicating "poor fit") with sample sizes greater than 200. Moreover, because it is a relative measure of fit, its value can only be assessed by comparing multiple models. Unlike the χ^2 index, the TLI, CFI, and SRMR indices are absolute measures of model fit. In general, TLI and CFI values greater than .90 are considered "good"; SRMR values of .05 or less are considered "good."

Using these criteria, the one-factor baseline model exhibited acceptable fit with the data. The results were as follows: χ^2 (df=5) = 1860.622, p.< 0001, CFI = .902, TLI = .802, and SRMR = .06. In addition, all of the item factor loadings were statistically significant (see Table 1). Taken together, these results suggest that a one-factor model provides reasonable fit to the pooled sample.

During the second step, we estimated the fit of this one-factor model for each group. Measurement equivalence was assessed by constraining the factor loadings to be equal across the two groups. From a measurement perspective, to the extent that the "multiple groups" model provides similar fit to the "baseline model," one can conclude that the ratings are equivalent across groups.

The multiple groups model also exhibited acceptable fit with the data. The results were as follows: χ^2 (df=19) = 2431.159, p.< 0001, CFI = .874, TLI = .867, and SRMR = .07. Although the difference in χ^2 values was significant (570.537, df=14), this was not unexpected, given the extremely large sample size. More convincing are the CFI, TLI, and SRMR values, which were virtually unchanged from the baseline model. Moreover, the item factor loadings were again found to be statistically significant (see Table 2).

Taken together, these results suggest that both groups were responding to the scale with a common frame of reference. This allows us to proceed with the subsequent hypothesis tests (described below) using a common criterion measure.

HYPOTHESIS TESTS

The study's primary hypothesis was tested using hierarchical multiple regression (see Table 3). The pilots' rank and number of flight hours were entered in Step 1. Together, they explained only .4% of the variance in pilots' reactions to their check ride. This was not unexpected, because the design and administration of check rides – both maneuver validations and LOEs – are strongly proscribed by the FAA, and are applied equally regardless of the pilots' rank or experience.

Pilots' reactions to CRM training and their recognition of the balance between training and checking were entered in Step 2. Together, these two factors explained both statistically and practically significant amounts of incremental variance (38.7%) in the pilots' reactions to check rides. Although both factors were statistically significant, the effect of training/evaluation balance

(β = .47), was substantially larger than the pilots' experiences in CRM training (β = .26).

We then tested the difference between these two partial regression coefficients (Cohen & Cohen, 1983), and found that they were significantly different from one another $t(6946)=11.93$, $p<.01$. This suggests that training/evaluation balance has a significantly larger effect on pilots' reactions to check rides than their experiences in CRM training. Although we had made no hypothesis as to which factor would exert a stronger effect, these results make a great deal of sense. Our personal experience suggests that many pilots do not perceive the check ride to be part of the training proper. As a result, those pilots who recognize the necessity of coupling training with evaluation might be expected to have more positive reactions toward their check ride.

The type of check ride was entered in Step 3. After controlling for the covariates that were entered in steps 1–2, the type of check ride was unrelated to pilots' reactions to their check ride. Specifically, the type of check ride explained only .3% incremental variance in the pilots' reactions to their check ride. Moreover, the estimated marginal means for pilots who received a maneuver validation and LOE were 3.77 and 3.86, respectively. These are virtually identical to the uncorrected means, which were 3.73 and 3.90, respectively. Taken together, the results suggest that airline pilots perceive both types of check rides as being equally effective.

CONCLUSIONS

Despite having a large sample, a representative pool of respondents, a high degree of statistical power, and reliable scales, the data suggest that there is no best way to assess pilot performance at the end of training. Rather, both the maneuver validation and Line Operational Evaluation appear to be equally effective.

One interesting finding was the strong, positive effects of the training/evaluation balance and experiences in CRM. These findings are extremely important, because carriers can positively influence these attitudes. For example, carriers can foster a positive training/evaluation balance by conducting employee satisfaction surveys, and by ensuring that pilots are not advanced to the check ride until they are ready. Carriers can also foster positive attitudes toward CRM training by integrating CRM skills training throughout the entire curriculum (Baker et al., 2002; Beaubien & Baker, 2002).

As with every study, this one has its limitations. First, we recognize that utility reactions are only one measure of training effectiveness. However, because trainees are the consumers of training, they are a valuable source of information regarding its effectiveness. Moreover, research by Kraiger and colleagues (1993) has shown that trainee satisfaction is an important outcome of training, while Alliger and colleagues (1997) have shown that utility reactions are positively correlated with learning and training transfer. Therefore, the more that trainees are satisfied with their training and find it useful, the more likely that training will be effective.

Second, we recognize that these results need to be replicated with other criterion measures – such as learning or behavioral transfer – to assess their generalizability. Because this is only a first look at the relative effectiveness of different forms of check rides, we caution the reader against making unwarranted generalizations.

In light of these limitations, we refrain from making recommendations for practice. However, we do recommend that researchers continue to explore the effectiveness of techniques for checking pilot performance at the end of training. Because the check ride is an essential component of pilot training, it plays a major role in ensuring the margin of safety. We believe that this line of research deserves additional attention by both researchers and practitioners.

REFERENCES

Alliger, G. M., Tannenbaum, S. I., Bennett, W. J., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology, 50*, 341-358.

Baker, D., Beaubien, J. M., & Mulqueen, C. (2002). *Airline pilot training survey: Final report*. Washington, DC: American Institutes for Research.

Beaubien, J. M., & Baker, D.P. (2002). Airline pilots' perceptions of and experiences in Crew Resource Management (CRM) training. In *Proceedings of the 2002 Society of Automotive Engineers' World Aviation Congress and Display*. Washington, DC: Society of Automotive Engineers.

Boehm-Davis, D. A., Holt, R. W., & Hansberger. (1999). Pilot abilities and performance. In R. S. Jensen (Ed.), *Proceedings of the 9th International Symposium on Aviation Psychology*. Columbus, OH: The Ohio State University Press.

Bollen, K. A., & Long, J. S. (Eds). (1993). *Testing structural equation models*. Thousand Oaks, CA: Sage.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum

Cannon-Bowers, J. A., Prince, C., Salas, E., Owens, J. M., Morgan, B. B., & Gonos, G. H. (1989). Determining aircrew coordination training effectiveness. In *Proceedings of the 11th Annual Meeting of Interservice/Industry Training Simulation and Education*

Conference (pp. 128-136). Arlington, VA: National Defense Industrial Association.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.

Federal Aviation Administration. (1991). *Advanced qualification program* (Advisory Circular 120-54). Washington, DC: U.S. Department of Transportation.

Federal Aviation Administration. (1990a). *Line operational simulations: Line oriented flight training, special purpose operational training, line operational evaluation* (Advisory Circular 120-35B). Washington, DC: U.S. Department of Transportation.

Federal Aviation Administration. (1990b). *Special Federal Aviation Regulation 58 – Advanced Qualification Program. Federal Register, Vol. 55, No. 91, Rules and Regulations* (pp. 40262-40278). Washington, DC: National Archives and Records Administration.

Helmreich, R. L. and H. C. Foushee (1993). Why crew resource management? Empirical and theoretical bases of human factors training in aviation. In E. L. Weiner, B. G. Kanki and R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 3-45). San Diego, CA: Academic Press.

Helmreich, R. L., Weiner, E. L., & Kanki, B. G. (1993). The future of crew resource management in the cockpit and elsewhere. In E. L. Weiner, B. G. Kanki & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 479-501). San Diego, CA: Academic Press.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78*, 311-328.

National Transportation Safety Board. (1994). *A review of flightcrew-involved, major accidents of U.S. air carriers, 1978 through 1990*. Safety Study NTSB / SS-94 / 01, Notation 6241. Washington, DC: Author.

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, 43*, 641-674.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd edition). New York: HarperCollins.

Yammarino, F. J., Skinner, S. J., & Childers, T. L. (1991). Understanding mail survey response behavior: A meta-analysis. *Public Opinion Quarterly, 55*, 613-63.

ACKNOWLEDGMENTS

This research was supported through grant #99-G-0048 from the FAA's Office of the Chief Scientific and Technical Advisor for Human Factors, Dr. David P. Baker, Principal Investigator.

The views expressed in this paper are solely those of the authors, and do not necessarily reflect those of the FAA.

The authors would like to thank Dr. Amy Nicole Salvaggio for her assistance in conducting some of the statistical analyses.

CONTACT INFORMATION

David P. Baker, Ph.D.
Principal Research Scientist
American Institutes for Research
1000 Thomas Jefferson Street, NW
Washington, DC 20007-3835
202-342-5036 (phone)
202-342-5033 (fax)
dbaker@air.org (e-mail)

J. Matthew Beaubien, Ph.D.
Senior Research Scientist
American Institutes for Research
1000 Thomas Jefferson Street, NW
Washington, DC 20007-3835
202-342-5133 (phone)
202-342-5033 (fax)
jbeaubien@air.org (e-mail)

Gonzalo Ferro, M.A.
Research Associate
American Institutes for Research
1000 Thomas Jefferson Street, NW
Washington, DC 20007-3835
202-298-2658 (phone)
202-342-5033 (fax)
gferro@air.org (e-mail)

Table 1.
Factor Loadings for the Check Ride Scale (Baseline Model)

Item #	Regression Weight	Standard Error	Z-Score	Standardized Estimate
1	1.000			.827
2	1.023	.017	61.123	.846
3	.868	.014	62.613	.717
4	.666	.012	53.858	.550
5	.963	.016	61.771	.796

Note: n = 8067

Table 2.
Factor Loadings for the Check Ride Scale (Two-Group Model)

Item #	Group	Regression Weight	Standard Error	Z-Score	Standardized Estimate
1	Maneuver Validation	1.00			.822
	Line-Oriented Evaluation	1.00			.822
2	Maneuver Validation	1.024	.016	62.652	.842
	Line-Oriented Evaluation	1.024	.016	62.652	.842
3	Maneuver Validation	.871	.014	63.830	.716
	Line-Oriented Evaluation	.871	.014	63.830	.716
4	Maneuver Validation	.668	.012	54.552	.549
	Line-Oriented Evaluation	.668	.012	54.552	.549
5	Maneuver Validation	.960	.015	62.628	.789
	Line-Oriented Evaluation	.960	.015	62.628	.789

Note: n = 4008 (MV) and 4059 (LOE)

Table 3.
Dependent Variable: Pilots' Reactions to Their Check Rides

Step	Variable Name	b	Std. Error	β	Sig.	R^2 Change	Power
1	Rank (Capt. vs. First Officer)	.075	.024	.047	.002	.004	1.00
	Number of Flight Hours	-.008	.006	-.022	.136		
2	Reactions to CRM Training	.231	.009	.261	.000	.387	1.00
	Training/Checking Balance	.462	.010	.467	.000		
3	Check Ride (MV vs. LOE)	.095	.015	.059	.000	.003	1.00

Note 1: Power was estimated at n = 6952, medium effect size (R^2 = .15), 5 predictors, and α = .05
Note 2: Actual R^2 = .394 (Adjusted R^2 = .393)