Journal of Aviation/Aerospace Education & Research

Volume 25
Number 2 *JAAER Spring 2016*

Article 2

Spring 2016

Judging Airline Pilots' Performance With and Without an Assessment Model: A Comparison Study of the Scoring of Raters From Two Different Airlines

David Weber
Griffith University, david.weber@originenergy.com.au

Introduction

The effectiveness of Crew Resource Management (CRM) programs has been evaluated previously (O'Connor, Jones, McCauley, & Buttrey, 2012), citing (in part) a lack of evidence to support the view that CRM training is having a clear effect on the mishap rate. One example that raises such questions and concerns is the fatal accident of Crossair flight 3579, which involved a highly experienced airline crew that was assessed and deemed capable of performing its duty only one month before the occurrence of the accident (Aircraft Accident Investigation Bureau, [AAIB], 2001).

These findings are consistent with other research results (e.g., Salas, Wilson, Burke, & Wightman, 2006). One of the specific problems in this is how ergonomics, (both researchers and practitioners), can determine whether the behavioral categories typically used in CRM training and assessment are linked to actual safety-critical behavior in cockpits. Furthermore, it needs to be questioned whether the very use of such categories itself influences or even determines examiners' construction of what occurs in that cockpit.

A better understanding of airline pilot performance assessment is the subject of ongoing research (e.g., Brannick, Prince, & Salas, 2002; Deaton et al., 2007; Dismukes, McDonnell, & Jobe, 2009; Holt, Hansberger, & Boehm-Davis, 2002; Mavin & Roth, 2014; Mulqueen, Baker, & Dismukes, 2002; Salas, Bowers, & Prince, 1998; Thomas, 2004). Various models are described in the literature to assess the performance of airline pilots. Some examples are the use of behavioral markers (Helmreich & Foushee, 1993), Line Operational Evaluations (LOE) (e.g., Goldsmith & Johnson, 2002; Hamman & Holt, 1997), NOTECHS (Flin et al., 2003; O'Connor et al., 2002), or the Model for Assessing a Pilot's Performance (MAPP) (Mavin & Roth, 2014). However, the use of models is hardly standardized among airlines (Mavin, Roth, & Dekker,

2013), particularly for the assessment of pilots' non-technical skills (i.e., situation awareness, communication).

The development of assessment models and grade sheets has proven difficult (Goldsmith & Johnson, 2002). Large efforts have been made to categorize the variety of crew behavior into a fine set of behavioral markers (e.g., Dutra, Norman, Malone, McDougall, & Edens, 1995; Flin & Martin, 2001). Yet doubt has been expressed about breaking complex performance into smaller, measurable components (Mavin & Dall'Alba, 2011; Mavin & Roth, 2014).

It is a common trend to measure the value of an assessment model in terms of its (inter-rater) reliability and validity measures, which build on assessment scores of various independent raters (e.g., Hamman & Holt, 1997; Holt, Johnson, & Goldsmith, 1997; O'Connor et al., 2002). However, studies have revealed large variance and low inter-rater reliability in the scoring of pilot performance (i.e., Mavin et al., 2013; Roth & Mavin, 2013; Roth, Mavin, & Munro, 2014). Pursuing the goal of agreement across raters, several strategies are widely used to reduce such rater disagreement (Baker & Dismukes, 2002; Bernardin & Buckley, 1981; Goldsmith & Johnson, 2002; Woehr & Huffcutt, 1994). These strategies aspire to increase assessors' observational skills (behavioral-observation training), to teach assessors to avoid commonly made mistakes (rater-error training), to familiarize raters with the assessment scale (performance-dimension training), and to instruct raters on performance assessment according to a standard set by experts (frame-of-reference training). Hence, efforts to increase raters' agreement largely focus on the assessors themselves.

In contrast, the constructs, concepts, and categories underlying assessment models and human factors are hardly questioned any longer (e.g., Dekker, Nyce, van Winsen, & Henriqson, 2010). The literature lacks an investigation into the consequences of using a model in the

assessment of complex pilot performance. It remains unclear whether the use of a model has an influence on the raters' scoring in the assessment process. The scores of raters who use a model have not yet been contrasted with the scores of raters who assessed performance without a model. The need of the present study is reflected in this lack of statistical examinations into the scoring of performance in the airline aviation industry and research.

The aim of this study was to draw a comparison of assessment scores provided by raters from two different airlines when judging the performance of peers (captain and first officer) in multiple video-scenarios. Assessors judged pilot performance in pairs of equal rank (first officers, captains, and flight examiners). One participating airline used an assessment model, whereas the other one was not given a model. In the light of this contrast, we examined differences in the scoring of assessors-pairs. Further exploratory examinations were conducted, pertaining, for example, to assessors' critical observation that passengers may have been evacuated from the aircraft to the side of a spinning engine.

Methods

Participants

The present study involved two airlines, Airline A and Airline B, which both conduct passenger flights in the Southern hemisphere on the ATR-72 and Dash-8 aircraft, respectively. The data collection of Airline A preceded the data collection of Airline B. In total, $N = 36$ airline pilots (subsequently referred to as assessor/s, assessor-pilot/s, or [assessor] pair/s) participated, $n = 18$ of each airline: 6 first officers, 6 captains, and 6 flight examiners. All assessors held a current license and operated the respective aircraft according to the rank they held when they participated in the present study. Participants were randomly chosen among those with free slots during the one-week data collection period of each airline. Separately for

Airline A and B, the assessors were randomly grouped into pairs of equal rank: first officer pairs (subsequently referred to as FO), captain pairs (CAP), and flight examiner pairs (FE). In total, there were 9 pairs per airline and 3 pairs per rank. The demographics of the participating assessors (age, total flight hours, and years as commercial pilot) are outlined in the Results section (see *Demographics of the assessors*; Table 5; Figure 4). The reason to conduct assessments in pairs is mirrored in the encouragement of producing oral assessment protocols, which have been deemed more natural than the use of *think-aloud protocols* by individual assessors (e.g. Ericsson & Simon, 1993).

Design

The analysis involved an observational 2 x 2 x 3 x 3 between groups four factorial design. The independent variables (subsequently referred to as IV) were *Airline*, *Pilot assessed*, *Rank*, and *Scenario* (variables are italicized and first letter capitalized throughout the present study). There were two levels of the IV *Airline:* Airline A, Airline B (nominal level of measurement); two levels of *Pilot assessed:* captain, first officer (nominal); three levels of *Rank:* FO, CAP, FE (treated ordinal, because the FO were deemed the least experienced, whereas the FE the most experienced assessors); and three levels of *Scenario:* Scenario 1, Scenario 2, Scenario 3 (nominal).

Two additional variables were added in the course of the analysis: the IV *Groups* and the dependent variable (DV) *Spinning engine*. *Groups*, on the one hand, was necessary to draw a comparison of the scores provided to each pilot assessed and scenario by the raters of the two airlines. It was derived from the variables *Airline*, *Pilot assessed*, and *Scenario*. There were twelve levels of *Groups* (nominal): A1-F1 and A2-F2 (Table 3). *Spinning engine*, on the other hand, allowed us to analyze whether the assessor-pairs noticed that the engine was still spinning

when the pilot crew evacuated the passengers to this side in Scenario 3. The *Spinning engine* was dichotomous: noticed, not noticed (nominal).

The DV *Scores* were treated as a continuous (interval) because it represents a Likert-type scale, assuming equal distances on each item (i.e., Norman, 2010; van Alphen, Halfens, Hasman, & Imbos, 1994). The scores were ranked as follows: 1 (*unsatisfactory performance*), 2 (*minimum standard*), 3 (*satisfactory*), 4 (*good standard*), and 5 (*very good standard*). This type of scoring represents the assessment scale both participating airlines use to assess the performance of their pilots. The analysis involved a total of 108 assessment scores: 2 airlines x 9 pairs x 3 scenarios x 2 pilots assessed.

The IV *Airline* involved a confound that could not be entirely removed. One group of assessors was given an assessment model and instructed how to use it, whereas the other group was not. Hence, the distinguishing factor between the treatments of the two groups was the airline. Strictly speaking, the resulting confound prevented to determine whether significant differences were due to the model used or the airline involved. In other words, differences in raters' scoring could either have resulted from the fact that the two airlines were involved in this research or that one airline used an assessment model to judge performance whereas the other did not. The results thus have to be interpreted with care. However, this confound was partly controlled by distinguishing between different ranks of assessors. Furthermore, all participating flight examiners were licensed and authorized by both their airline and the regulator to determine the level of pilots' performance and make decisions that may affect career options.

This authorization is independent of the model (or other assessment tools) applied, the training received, or the airline involved. Furthermore, the experience of the assessor-pilots of Airline A and B were equated and controlled by comparing the demographics (age, total flight

hours, years as commercial pilot) overall and in regards to specific ranks (see *Demographics of the assessors*).

Scenarios

Each pair assessed three videotaped scenarios, showing the performance of a captain and first officer flying as a crew in a certified full-motion flight simulator (abbreviations were only used to refer to the assessor pairs [FO, CAP, FE], yet not to refer to the pilots [captain and first officer] assessed in the scenarios). Both pilots in the scenarios wore company uniforms. Other employees assumed roles as the cabin crew and ATC controllers.

All scenarios were scripted in advance and recorded with three cameras from the position of the flight examiner (seated behind the pilots). The first camera captured the pilots from behind; the second and third cameras provided pertinent close-up pictures of relevant instrument displays. At several stages during the scenarios, relevant charts (such as approach plates) were superimposed.

In Scenario 1 (subsequently referred to as S1), a captain (pilot flying) and first officer conducted an instrument approach at daytime. Close to the airport, they became visual with the runway and initiated a circling approach to land against the wind from the opposite direction. During base turn, the aircraft encountered a rain cloud, which required the crew to conduct a missed approach. The captain initially turned the wrong way. However, he was immediately corrected by the first officer.

Scenario 2 (S2) showed a captain and first officer (pilot flying) performing an instrument approach by day. Poor weather hampered visual contact with the runway, which forced the crew to conduct a missed approach. Despite the low fuel status, the captain attempted to convince the first officer to try a second approach at the destination airport and only divert to the alternate

airport in case they would fail to land. The first officer was reluctant and expressed his intention to directly divert to the alternate airport.

In Scenario 3 (S3), a captain (pilot flying) and first officer conducted an approach to an airport during daylight. When the aircraft was already configured for landing (landing gear down, flaps set), the left engine caught fire. Both crew members followed their procedures to extinguish the fire and continued on the approach. Seconds before touching down on the runway, the first officer informed the cabin crew about the present situation. The captain landed the aircraft and ordered the passengers to evacuate on the runway after having received clearance from the tower controller.

The footage of the scenarios presented to the pairs of Airline A and B slightly differed. Whereas S3 was identical for the pairs of both airlines (aircraft: Dash-8; duration: 9.16 min), S1 and S2 of Airline B were reproductions of the respective scenarios of Airline A. The reason was that all assessors had to be familiar with the aircraft, cockpit, environment, airports, and approaches shown in the scenarios. Consequently, S1 of Airline A was filmed in an ATR-72 simulator (duration: 6.45 min), whereas its reproduction of Airline B was filmed in a Dash-8 simulator (7.54 min). The same applied to S2 (Airline A: ATR-72, 3.30 min; Airline B: Dash-8, 3.46 min). However, before and during the filming of the reproduced scenarios, the original footage was presented to the acting crew. The takes were repeated until the scenarios were highly similar in terms of their content, action plot, pilot performance, and conversation. The duration of the original and reproduced S1 differed by 69 seconds, whereas S2 by 16 seconds. This variation was unavoidable due to the different approach procedures of the destination airports in the scenarios. The authors are aware of the less than optimal footage presented to the

assessors. However, the high similarity of the original and reproduced scenarios justifies a comparison of the scoring between the assessors of Airline A and B.

Procedure

The study setting (Figure 1) was identical for both airlines. Each assessor pair (comprised of Assessor 1 and Assessor 2) was seated at a table in front of a large LCD TV screen, which they controlled via a computer with a mouse. The pairs were asked to keep all sheets in the designated area of work, which was recorded from above with a camera (CAM). CAM3 recorded the notes taken, the scores provided on the assessment sheets, assessors' pointing gestures, etc. CAM1 captured the assessors and the area of work, whereas CAM2 provided a closer view of the assessors.

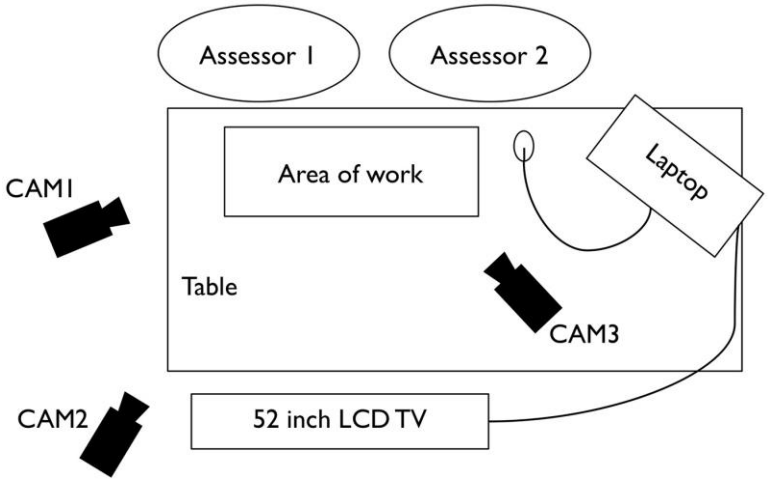

Figure 1. Study setting.

The only difference between the data collection of the two airlines was the written material provided to the pairs of Airline A and B during the assessment. The pilot pairs of Airline A were given two sheets from their company's training manual: the model used by Airline A to assess their pilots, and an assessment form (Table 1) derived from the model. The assessor pairs only received one sheet for each pilot assessed (one for the captain and one for the

first officer). The assessment form provided detailed descriptions of performance for each of the scores (1 through 5) and assessment categories ('Situation awareness,' 'Decision making,' 'Aircraft flown within tolerances,' 'Knowledge,' 'Management,' and 'Communication') of the model. In accordance with the literature (e.g., Mavin & Dall'Alba, 2011; Mavin & Roth, 2014; Mavin et al., 2013; Roth et al., 2014), the descriptions of aspects of performance were referred to as 'word pictures'.

Table 1

Excerpt of the Assessment Form, Including the Word Pictures of the Category 'Management'

	Scores				
	1	2	3	4	5
MN • Workload • Control • Cooperation • Threats & Errors	• Ineffective organisation of crew tasks. • Inability to control self or crew member performance. • Interaction was negligible, or disrupted team effectiveness. • Serious threats or errors not mitigated or managed.	• Inefficient organisation of crew tasks. • Controlled self or crew member actions, though with difficulty. • Interacted with crew member, but provided limited support. • Threats or errors not well mitigated or managed.	• Adequate organisation of crew tasks. • Controlled self or crew members performance; disagreements resolved. • Interacted with crew member. • Most threats managed; most errors trapped.	• Crew member tasks effectively organized. • Effective control of self or crew to achieve expected performance. • Considered other crew to improve team performance. • Threats identified and managed; errors trapped.	• Tasks organised so challenging aspects of flight appeared easy. • Effective control of self or crew, even in challenging situation. • Interaction with and consideration of crew maximized performance. • TEM well integrated.

Note. MN=Management, 1-5 = assessment scores (1 = *very poor performance*, 5 = *very good performance*).

In contrast, the pairs of Airline B were only given one assessment-sheet per scenario, which neither provided word pictures nor categories. It only listed the scores (1 through 5), separately for each pilot assessed, together with the identical descriptions of the scores that were used in the study of Airline A. All assessors of both airlines were given blank sheets to take as many notes as they wanted. Whereas participants of Airline B did not receive assessment training from the researchers (in addition to assessment training within their organization), assessors of Airline A were familiarized with the assessment model and form prior to participating in this study. The researchers provided PowerPoint presentations and discussion sessions to introduce all Airline A assessors (including first officers, captains, and flight examiners) to the model and assessment form in the course of the pilots' CRM training. Every pilot had used the model and assessment form to judge the performance of peers in at least three video scenarios other than the ones assessed in the present study. Results were compared and discussed. All assessors of Airline A thus had been exposed to the model, assessment form, and theoretical concepts (e.g. Situation awareness, Decision making) to a highly similar extent.

All pairs were given the same task: to watch, discuss, and assess the performance of the captain and first officer in each scenario. To encourage the verbalization of their reasoning, all participants assessed performance in pairs. All participants were informed that the aim of the study was to better understand the reasoning behind the assessment of pilot performance.

The pairs of both airlines were free to pause, replay, or go back in the video scenarios at any time and as often as they wanted. The decision of where to start their discussion and how to assess performance was left entirely to the assessors. Two researchers attended the assessment sessions. Following a fixed protocol, they only intervened when it was necessary to encourage participants to speak louder, to clarify a comment, or to provide assessment scores.

To motivate the raters express their thoughts, each pair of assessors only received one scoring sheet for each pilot assessed. This required each pair to come to an agreement about the score for each pilot (pairs were allowed to provide semi-scores, e.g. 4.5, if they could not come to an agreement). After their discussion, the pairs of Airline B circled one score for the captain and one score for the first officer. In contrast, the pairs of Airline A assessed one category ('Situation Awareness,' 'Decision making,' etc.) after the other. They circled specific word pictures on the assessment form and eventually decided about the pilot's score for each category. To draw a comparison between the scoring of the two airlines, we calculated an average overall score for each pilot and scenario assessed by the pairs of Airline A: the sum of the scores for each category, divided by the number of categories (6).

Data Analysis

After the data collection, a picture-in-picture video was created for each of the 18 assessment sessions, including the footage of all three cameras. The pairs' scores were collected in an Excel spreadsheet, separately for each scenario, pilot assessed, and airline. In the course of the assessment sessions, all $N = 36$ participants were asked about their demographics: *Age* (in years; interval), *Total flight hours* (interval), and *Years as commercial pilot* (CPL; interval). All variables involved in the analysis were derived from the videos, scores, and/or demographics. The data were analyzed using the Statistical Package for the Social Sciences (SPSS) software, Version 22. The data was manually entered and verified before the below specified statistical procedures and tests were performed.

Rather than applying multivariate procedures, we calculated univariate analyses due to the following reasons: first, the demographic variables pertained to individual assessors, whereas the scores related to the assessor pairs. We analyzed the age, total flight hours, and years flown

as commercial pilot separately from the analysis of the scores (see *Demographics of the assessors*). Second, the issue of the spinning engine only related to S3. We conducted a separate analysis of the spinning engine (see *The spinning engine in Scenario 3*). Third, we looked into the pass-fail rating independently of the scores because the former was partly derived from the latter (see *Pass-fail rating*).

Results

In the following, we present our findings in regards to the scores provided by the pairs of the two airlines (see *Analysis of the scores*). Subsequently, we outline further relevant and interesting findings (see *Further exploratory analyses*).

The authors are aware of the problem related to the small sample size and the relatively large number of tests employed in this study, which increased the likelihood of spuriously significant results. However, throughout the entire analysis, all *p*-values of post hoc contrasts were adjusted/corrected for multiple comparisons according to Bonferroni. If not stated otherwise, an alpha level of 5% and two-directional hypotheses/*p*-values were used.

Analysis of the Scores

Each assessor pair provided a score for the captain and the first officer assessed in each scenario (Table 2). The scores of Airline B assessor pairs were always equal or lower than those of Airline A. Only one of the CAP pairs of Airline B assessed the first officer in S1 with a higher score than the pairs of Airline A. However, except for the first officer assessed by the FE pairs in S2, there were always pairs of Airline B that provided the same scores as some of the pairs of Airline A.

Table 2

Scores Provided by the Assessor Pairs of Airline A and B to the Captain and First Officer Assessed in the Scenarios

		Airline A		Airline B	
Scenario	Assessor pair	Captain	First officer	Captain	First officer
S1	FO	3, 4	4	2, 3	3, 4
	CAP	2, 3	4	1, 2	2, 4, 5
	FE	2, 3	3, 4	1, 2	1, 2, 3
S2	FO	4	3, 5	3, 4	2, 4, 5
	CAP	2, 3, 4	4, 5	1, 2	1, 3, 4
	FE	3, 4	4, 5	1, 2, 3	1, 3
S3	FO	4, 5	4, 5	2, 4	2, 4
	CAP	3, 5	4, 5	1, 4, 5	1, 4, 5
	FE	2, 3, 4	2, 3, 4	1, 2	1, 2

Note. S1 = Scenario 1, S2 = Scenario 2, S3 = Scenario 3, FO = First officer assessor (pair), CAP = Captain assessor (pair), FE = Flight examiner assessor (pair), Captain/First officer = Pilots assessed.

The mean scores are shown in Figure 2, separately for each pilot assessed, airline, scenario, and the pairs' rank. The mean scores of the Airline B pairs were always lower than Airline A, which applies to both pilots assessed. On average, all pairs scored the first officer higher than the captain in S1 and S2. In S3, the captain and first officer received identical mean scores, except for the FO pairs of Airline A that scored the captain slightly higher (score 4.67) than the first officer (score 4.33).

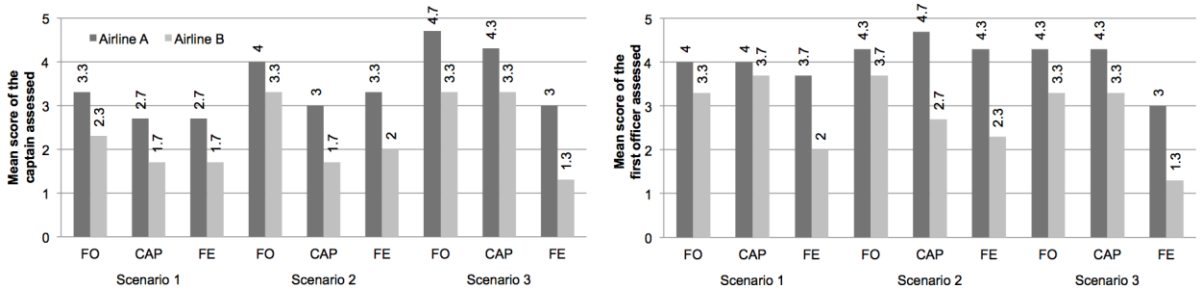

Figure 2. Mean scores provided to the captain and first officer assessed in each scenario by the assessor pairs of different rank and airline.

The aim of the present effort was to examine whether there is more agreement in terms of the scores if raters use an assessment model in the evaluation of pilot performance. Given that Airline A used a model whereas Airline B did not, we drew a comparison of the pairs' scoring between the two airlines (the authors are aware of the limitations to this approach, which will be addressed in the discussion). However, such a comparison is only valid when comparing the scores that the pairs of each airline provided to each pilot assessed and scenario. This, in turn, required a distinction between six groups (Table 3): A1/A2, B1/B2, C1/C2, D1/D2, E1/E2, and F1/F2. 'Group 1' (A1-F1) related to Airline A, whereas 'Group 2' (A2-F2) to Airline B.

Table 3

Dispersion Measures of the Scores of Each Group

	Airline A		Airline B	
	Captain	First officer	Captain	First officer
S1	A1: 2.89; 0.601; 2	D1: 3.89; 0.333; 1	A2: 1.89; 0.601; 2	D2: 3.00; 1.225; 4
S2	B1: 3.44; 0.726; 2	E1: 4.44; 0.726; 2	B2: 2.33; 1.000; 3	E2: 2.89; 1.364; 4
S3	C1: 4.00; 1.118; 3	F1: 3.89; 0.928; 3	C2: 2.67; 1.581; 4	F2: 2.67; 1.581; 4

Note. Group (e.g. A1): Mean (e.g. 2.89); SD (e.g. 0.601); Range (e.g. 2); S = Scenario.

An answer to the question about whether there is more agreement in terms of the scoring if assessors use a model could not be found by simply comparing the mean scores of the two airlines, because the scores related to three different scenarios and involved a captain and a first officer being assessed. Rather, the pairs' agreement in terms of the scores is reflected in the groups' dispersion measures (standard deviation [SD] and range of the scores) between each two groups (e.g. A1/A2). We thus compared the dispersion measures of each two groups (Table 3).

Each group's mean, SD, and range was based on the 9 scores that the 9 pairs (3 FO, 3 CAP, 3 FE) of the respective airline provided to each pilot assessed in each scenario. For instance, the value 2.89 (Table 3) reflects group A1's *mean* of the 9 scores provided to the captain assessed in S1 by the three FO, three CAP, and three FE pairs of Airline A.

First, we compared the *SD* of each two groups. We tested the hypothesis (H_1, one-tailed) that the mean SD of the scores is greater for the groups of Airline B (A2 to F2) than for Airline A (A1 to F1), because the assessor pairs of Airline B were not given an assessment model (H_0*:* $M_{\text{Airline A}} = M_{\text{Airline B}}$; H_1: $M_{\text{Airline A}} < M_{\text{Airline B}}$). A paired-samples *t*-test (comparing each two groups) showed that the mean SD of the scores of Airline B (M = 1.23, SD = 0.38) was significantly higher than those of Airline A (M = 0. 74, SD = .27), $t(5) = -3.78$, $p = .007$ (one-tailed), d = -1.54, 95 CI [-.82, -.16].

A (non-parametric) sign-test (H_0: $p = 0.5$; H_1: $p \neq 0.5$) fortified the result of the *t*-test: The SD of each two groups' scores were unequal between Airline A and B, $p = .032$ (one-tailed), with 0 negative differences (SD of Airline B-groups < SD Airline A), 5 positive differences (SD Airline B > SD Airline A), and 1 tie (SD Airline B = SD Airline A; which was the case for the groups A1/A2; Table 3).

Second, we tested the same hypothesis but in terms of the *ranges* of each two groups. The results of a paired-samples *t*-test showed that the mean range of the scores of the Airline B-groups (M = 3.50, SD = .837) were significantly higher than Airline A (M = 2.17, SD = .75), $t(5) = -3.16$, p = .013 (one-tailed), d = -1.29, 95 CI [-2.42, -.24].

A sign-test (testing the same hypothesis as above) showed the same result: the ranges of the scores were unequal for the groups of Airline A and B, p = .032 (one-tailed), with 0 negative differences (ranges of Airline B-groups < Airline A), 5 positive differences (ranges of Airline B > A), and 1 tie (ranges Airline B = A, which was the case for the groups A1/A2; Table 3).

The analysis of the dispersion measures revealed highly consistent results, even when using different measures (SD, range) and tests (parametric, non-parametric). The SD and ranges of the scores were different for the groups of the two airlines and greater for Airline B than for Airline A. Consequently, there was a smaller spread of the scores in Airline A. The results indicate that there is more agreement in terms of the scoring among pairs of Airline A, who used a model to judge pilots' performance.

In contrast to comparing the groups' dispersion measures, we further examined whether the groups of the two airlines (Table 3) significantly differ in terms of their actual scoring. Applying an independent samples *t*-test, the means of the scores of each two groups were compared (A1/A2, B1/B2 etc.; n = 9 per group). A t statistic not assuming homogeneity of variances was used when homogeneity of variances between the groups could not be assumed. The results showed significant differences of the mean scores between the groups A1/A2, $t(16) = 3.53$, p = .003, d = 1.66, 95 CI [0.40, 1.60], B1/B2, $t(16) = 2.70$, p = .016, d = 1.27, 95 CI [0.24, 1.99], and E1/E2, $t(16) = 3.02$, p = .008, d = 1.42, 95 CI [0.46, 2.65]. The differences between the other groups were not significant.

We calculated a 4-way analysis of variance (ANOVA) to examine whether the (DV) *Scores* was a function of the (IVs) *Scenario, Rank, Pilot assessed,* and *Airline*. The mean, SD, and sample size (n) for each level of the IVs are shown in Table 4.

Table 4

Sample Sizes (n), Means, and SD of the Scores in Relation to the IVs Rank, Pilot Assessed, Airline, and Scenario

IV	Levels of IV	n	Mean	SD
Rank	FO	36	3.67	0.926
	CAP	36	3.28	1.386
	FE	36	2.56	1.132
Pilot Assessed	Captain	54	2.87	1.182
	First Officer	54	3.46	1.239
Airline	Airline A	54	3.76	0.889
	Airline B	54	2.57	1.268
Scenario	Scenario 1	36	2.92	1.025
	Scenario 2	36	3.28	1.233
	Scenario 3	36	3.31	1.431

Note. IV = Independent variable, SD = Standard Deviation, FO = First officer assessor (pair), CAP = Captain assessor (pair), FE = Flight examiner assessor (pair).

The 4-way ANOVA revealed significant main effects for *Rank* [$F(2, 72) = 12.36$, $p < .001$, $\omega^2 = .127$] (ω^2 indicates that approximately 12.7% of the variation in the *Scores* was attributable to differences between the three groups of assessor pairs [FO, CAP, and FE]), *Pilot assessed* [$F(1, 72) = 10.24$, $p = .002$, $\omega^2 = .052$], and A*irline* [$F(1, 72) = 40.96$, $p < .001$, $\omega^2 = .223$], yet not for *Scenario*. However, we decided to keep the *Scenario* in the model due to its significant interactions with *Rank* [$F(4, 72) = 2.58$, $p = .044$, $\omega^2 = .035$] (Figure 3a) and *Pilot assessed* [$F(2, 72) = 3.25$, $p = .045$, $\omega^2 = .025$] (Figure 3b). (Homogeneous variances between

the two airlines in terms of the scores could not be assumed due to the differences found in the dispersion measures and the significant result of the test for homogeneity of variance [*Levene* $F(35, 72) = 2.37$, $p = .001$]. However, a Box-Cox transformation revealed that there is no better transformation of the DV ($\lambda = 1$). We deemed this the best model and build the findings on the original, non-transformed data.) These interactions contributed to additional 14% of variance explained ($r^2_{Scenario\ included} = .596$; $r^2_{Scenario\ excluded} = .456$). All other interactions were not significant. Yet due to its central role, we provide interaction plots for *Airline* and *Scenario* (Figure 3d), *Airline* and *Rank* (Figure 3e), as well as *Airline* and *Pilot assessed* (Figure 3f). For the sake of completeness, we also show the interaction of *Rank* and *Pilot assessed* (Figure 3c).

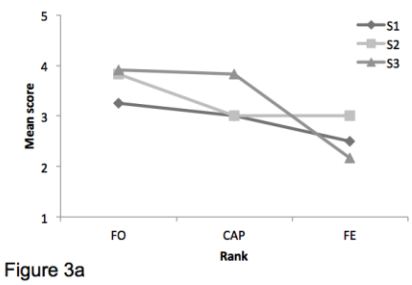

Figure 3a

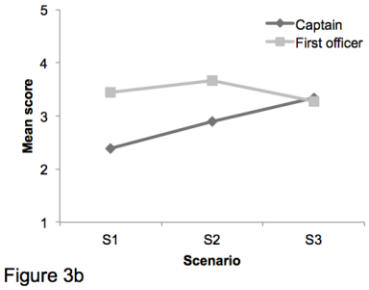

Figure 3b

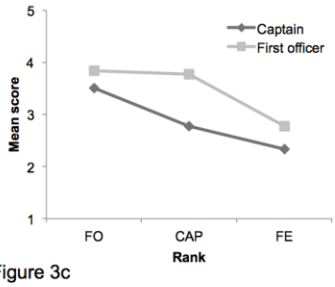

Figure 3c

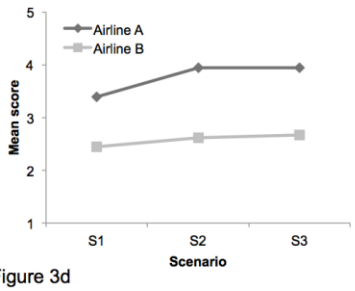

Figure 3d

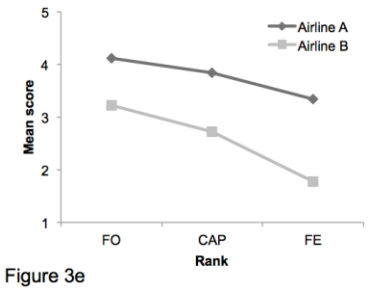

Figure 3e

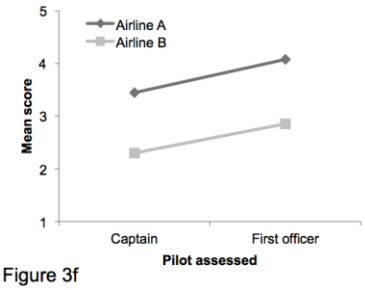

Figure 3f

Figure 3. Interactions of *Rank* and *Scenario* (3a), *Scenario* and *Pilot assessed* (3b), *Rank* and *Pilot assessed* (3c), *Scenario* and *Airline* (3d), *Rank* and *Airline* (3e), and *Pilot assessed* and *Airline* (3f).

Post hoc contrasts were calculated to provide pairwise comparisons (all *p*-values were corrected according to Bonferroni). The results indicated that the mean score (Table 4) provided to the first officer assessed was significantly higher than the mean score to the captain, $p = .002$, 95 CI [.22, .96]. On average, the pairs of Airline A provided significantly higher scores than the pairs of Airline B, $p < .001$, 95 CI [.82, 1.55]. The mean score of the FO pairs was significantly higher than the mean score of the FE pairs, $p < .001$, 95 CI [.55, 1.67]; also the CAP pairs scored significantly higher than the FE pairs, $p = .006$, 95 CI [.17, 1.28]; however, no significant difference was found between the scoring of the FO and CAP pairs. These findings were congruent with the correlation of the *Scores* and *Rank* [$r_s(108) = -.37$, $p<.001$]: the higher the pairs' rank (FE>CAP>FO), the lower the assessment scores provided.

The critical reader might question the interpretation of main effects when the respective IVs are involved in significant interactions. We calculated a 1-way ANOVA for each IV (in relation to the DV *Scores*), which showed identical significant results as in the more complex model (4-way ANOVA) outlined above: *Rank* [$F(2, 105) = 8.46$, $p < .001$], *Pilot assessed* [$F(1, 106) = 6.46$, $p = .012$], *Airline* [$F(1, 106) = 31.64$, $p < .001$], whereas *Scenario* again was not significant. The post hoc pairwise comparisons (Bonferroni corrected) were also identical.

Further Exploratory Analyses

In the following, we outline further relevant and interesting findings pertaining to the demographics of the participants, the important issue of a spinning engine in S3, and an examination of the pairs' pass-fail ratings.

Demographics of the assessors. The participants' ($N = 36$) mean age, total flight hours, and years as CPL are shown in Table 5 and visualized in Figure 4, separately for each airline and rank. Independent samples *t*-tests were conducted to compare whether the assessors of Airline A

and B significantly differed in terms of their demographic background. Comparisons were made between all assessors (n = 18 per airline), as well as between each individual rank (FO, CAP, FE; n = 6 per rank and airline). If equal variances could not be assumed (significant Levene's test), the findings for equal variances not assumed were reported, which is subsequently marked with an asterisk (*).

Table 5

Mean and SD of the Age, Total Flight Hours, and Years as CPL for Each Rank and Airline

	Rank	Age (years)	Total flight hours	Years as CPL
Airline A	FO	30.67 *(3.33)*	4950.00 *(2012.71)*	11.67 *(4.08)*
	CAP	45.33 *(6.68)*	15376.67 *(6096.10)*	24.17 *(6.68)*
	FE	49.17 *(6.21)*	14250.00 *(4906.22)*	25.33 *(5.92)*
	All assessors	41.72 *(9.75)*	11525.56 *(6505.07)*	20.39 *(8.30)*
Airline B	FO	26.50 *(3.21)*	3116.67 *(2835.08)*	6.00 *(4.29)*
	CAP	31.33 *(4.76)*	4800.00 *(1164.47)*	9.83 *(2.32)*
	FE	42.67 *(9.07)*	12050.00 *(4289.41)*	20.25 *(8.35)*
	All assessors	33.50 *(9.08)*	6655.56 *(4907.16)*	12.03 *(8.12)*

Note. Mean *(SD)*. SD = Standard Deviation, FO = First officer assessor (pair), CAP = Captain assessor (pair), FE = Flight examiner assessor (pair), CPL = Commercial pilot license (meaning: years as a commercial pilot).

The mean *age* of all assessors of Airline A (Table 5) was significantly higher than the mean age of the assessors of Airline B, $t(34) = 2.62$, $p = .013$, $d = 0.87$, 95 CI [1.84, 14.61]. The mean age of Airline A was also significantly higher than of Airline B in terms of the CAP assessors, $t(10) = 4.18$, $p = .002$, $d = 2.41$, 95 CI [6.54, 21.46]. In contrast, the mean age difference between Airline A and B was neither significant for the FE nor for the FO assessors.

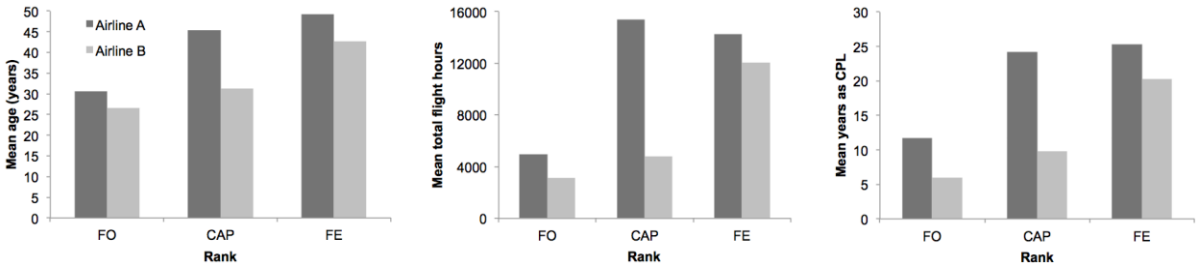

Figure 4. The participants' mean age, total flight hours, and years as CPL for each rank and airline.

The mean *total flight hours* of all assessors of Airline A (Table 5) was significantly higher than those of Airline B, $t(34) = 2.54$, $p = .016$, $d = .85$, 95 CI [966.88, 8773.12]. The mean total flight hours of Airline A was also significantly higher than those of Airline B in terms of the CAP assessors, $t(5.36) = 4.17$, $p = .007$*, $d = 2.41$, 95 CI [4194.38, 16958.96]. However, the difference of the mean total flight hours was neither significant in terms of the FE nor FO assessors.

The mean *yeas as CPL* of all assessors of Airline A (Table 5) was significantly higher than those of Airline B, $t(34) = 3.06$, $p = .004$, $d = 1.02$, 95 CI [2.80, 13.92]. The mean years as CPL of Airline A was also significantly higher than those of Airline B in terms of the CAP assessors, $t(6.19) = 4.97$, $p = .002$*, $d = 2.87$, 95 CI [7.33, 21.34], as well as the FO assessors, $t(10) = 2.34$, $p = .041$, $d = 1.35$, 95 CI [0.28, 11.05]. However, no significant difference of the mean years as CPL was found in terms of the FE assessors.

The spinning engine in Scenario 3. During the assessment of S3, some of the pairs noticed that an engine was still spinning when the crew ordered the passengers to evacuate the aircraft to this side. (Evacuating the passengers into a spinning engine was neither planned nor scripted in the scenario. It occurred during filming and was only noticed after the completion of the production. The authors believe that the unplanned nature of this issue contributed to the

authenticity of the scenario, which was the reason to not exclude S3 from the analysis.) In the discussion after the assessment, some of the pairs noted that the issue of the spinning engine was crucial and had strongly decreased the scores provided to both pilots assessed. We focused on S3 to provide an overview of the pairs who noticed the spinning engine or not, and analyzed the issue of the spinning engine in regards to the scores provided to the pilots.

The spinning engine was noticed by 4 out of 18 assessor pairs: one FE pair of Airline A, whereas two FE and one FO pair of Airline B. The FE pair of Airline A and FO pair of Airline B gave each pilot a score 2 for their performance in S3. The other two FE pairs of Airline B scored the captain and first officer a 1. All pairs who noticed the spinning engine failed both crew members. There was a significant difference between the mean scores of the pairs that noticed (mean score 1.50) vs. did not notice (mean score 3.82) the spinning engine, $t(34) = 5.47$, $p < .001$, $d = 2.19$, 95 CI [1.46, 3.18] ($n = 36$; only including S3).

Pass-fail rating. The pass-fail ratings came about differently for the two airlines: Airline A followed their company's fail-criterion, which resulted when the pilot was scored one rating of 1 or three ratings of 2 on any of the six assessment categories. In contrast, the pairs of Airline B could freely decide whether they passed or failed each pilot, due to not being given an assessment model. It was worth comparing the pass-fail ratings independently of the scores. An overview of the number of pass-fail ratings is given in Table 6, separately for each pilot assessed, scenario, and airline.

Table 6

Number of Pass-Fail Ratings for Each Airline, Scenario, and Pilot Assessed

			Pilot assessed		
			Captain	First officer	Total
Airline A	Scenario 1	Fail	5	1	6
		Pass	4	8	12
	Scenario 2	Fail	2	0	2
		Pass	7	9	16
	Scenario 3	Fail	1	1	2
		Pass	8	8	16
	All scenarios	Fail	8	2	10
		Pass	19	25	44
Airline B	Scenario 1	Fail	8	3	11
		Pass	1	6	7
	Scenario 2	Fail	5	3	8
		Pass	4	6	10
	Scenario 3	Fail	5	5	10
		Pass	4	4	8
	All scenarios	Fail	18	11	29
		Pass	9	16	25

The analysis of the number of pass-fail ratings revealed differences between the two airlines. Overall, the assessor pairs of Airline B failed the captain more often (18 times; Table 6) than the assessors of Airline A (8x). The same applied to the first officer (11x vs. 2x). In contrast, the assessors of Airline A passed both crewmembers more often (44x vs. 25x), which was also the case for each scenario assessed.

The comparison of the pass-fail ratings of all scenarios (Table 6, column 'total,' rows 'all scenarios') showed a significant difference between the two airlines, $p < .001$ (Pearson Chi-Square, $N = 108$). The results indicate that the airline that was not given a model to assess performance (Airline B) was much stricter in their assessment and failed both pilots more often than the pairs that used an assessment model (Airline A).

Discussion

This study was designed to examine differences in the scoring between assessors who used (Airline A) versus did not use (Airline B) an assessment model to judge pilots' performance. The results showed that the pairs of Airline B scored lower and failed the pilots more often than the assessors of Airline A. Furthermore, there was a larger spread of the scores in Airline B. In contrast to the first officer assessed, the pairs of the two airlines differed more extensively in their judgments of the captain, who also received the lower scores. It remains unclear whether this was due to poorer performance or, for instance, the higher level of responsibility associated with this role.

The present study is limited in its design due to a confound about whether the differences in the scoring are the result of the airlines involved or the use of an assessment model. In order to address this possibly confounding effect, various variables have been controlled in terms of the participating pilots' training and experience across the two airlines. Future research in this area may incorporate a more rigorous design, involving highly identical conditions across participants and integrating a larger number of pilots who operate the same type of aircraft and work for the same airline. This would standardize the amount of participants' CRM and assessment training among equal ranks. Allocating individual assessor-pilots to various groups that receive different treatments (e.g., not using a model, using Model A, using Model B, etc.)

will allow examining if variation in assessors' scoring is related to any particular treatment. The present study is partly limited in this regard due to accessibility to, and availability of, a larger number of airline professionals who all operate the same type of aircraft for the same airline, and with an equal amount of training and experience.

Rather than making a general statement of (inter-rater) reliability and validity (psychometrics), however, the authors express their concerns about the influence of an assessment model in the evaluation of pilot performance. The number of participants may be limited in this effort, yet the scores (as outlined in Table 2) clearly indicate differences, even within groups of the same airline and rank.

The critical reader might challenge whether differences in the scoring are the result of unequal levels of experience between Airline A and B. However, this reasonable doubt can largely be dispelled by assessors' similar demographic background. Namely, the flight examiners of Airline A and B did not significantly differ in terms of their age, flight hours, and years as CPL. The first officers solely differed in regards to their years as CPL. Only the captain assessors of Airline A were older, had logged more flight hours, and held their CPL for more years than their colleagues of Airline B.

However, it goes without saying that the experience of the participating pilots and their prior training cannot entirely be verified and validated for possible reasons of failure. Both have to be inferred. This too, however, was controlled in this study by thoroughly examining assessors' demographic background and by providing equal assessor training to each assessor group and airline. Yet pilots' experience and exposure to training may always slightly vary, for instance due to working in and for different airlines throughout their careers, having operated different types of aircraft, and variation across company procedures.

The training of the pilots may have slightly varied across the two airlines, yet all pilots, and notably the participating flight examiners, were highly experienced (Table 5). They had been nominated by their respective airline to perform their duties as examiners, independently of the type of aircraft they operate, the routes they fly, the hard limits set on performance standards, or standardization of experience and training across airlines. Differences in the very scoring of flight examiners between the two airlines and particularly *within* the same airline support the authors' concerns. Scores largely differ, namely the scores provided by the flight examiners. In fact, no pilot (captain or first officer assessed) received one particular score across all flight examiner pairs (Table 2). Flight examiners of Airline A and B differ in their scoring even when witnessing Scenario 3, which was identical for both airlines.

All pairs who noticed the spinning engine in S3 provided low scores and failed both crew members. However, it remains unclear why the spinning engine was not picked up by the other pairs and was observed by a larger number of pairs of Airline B. A critical issue was only noticed by a small number of assessor pairs, and by an even smaller number of pairs who did use an assessment model and word pictures. The use of a model apparently did not enhance assessors' observation of safety-critical crew performance. A critical observation resulted in substantial variation in assessors' scoring. This requires questioning whether an assessment model might better increase agreement among raters, if it directed assessors' attention to critical pilot behavior. Future research is required to examine whether the model used in the present study and other models are capable of supporting assessors' discovering of critical performance issues.

In accordance with behavioral-observation training (Woehr & Huffcutt, 1994), assessors have to be encouraged to watch and discuss scenarios in small sequences. It was this technique

that seemed to have enabled the pairs who noticed the spinning engine to discover hidden details. However, assessors hardly have unlimited time and the benefit of repeatedly watching the performance of pilots during their initial flight and recurrency training. Moreover, assessments do not only depend on assessors' observations, yet also on the way observations are *evaluated* and *weighted* (Borman, 1978). Consequently, different interpretations of complex pilot performance may remain likely.

When asked after the assessment of the three scenarios, the pairs of Airline A noted that the model and word pictures supported and matched what they referred to as their personal "gut feeling" about each pilot's performance. The pairs of Airline B pilots expressed their "gut feeling" too, despite not having used an assessment model. However, the pairs of the two airlines largely arrived at different conclusions about performance. The scoring not only differed between the pairs of Airline A and B, but also between pairs of the same airline and rank (Table 2). Rather than questioning the accuracy and adequacy of Airline A's performance review, there is the need for qualitative examinations into assessors' reasoning behind the scoring, particularly in regards to identical scores when using a model or not. Studies have to further investigate why scores differ, particularly if assessors all use the same assessment model and word pictures.

The confound in the design of this study prevents the generalization of the results to the larger population. However, the present investigation into a highly representative sample of airline professionals has outlined the importance of examining the influence of an assessment model when aspiring to better understand variation in the scoring of airline pilot performance.

References

Aircraft Accident Investigation Bureau (AAIB). (2001). Final report No. 1793, HB-IXM, Flight CRX 3597. Retrieved from http://www.sust.admin.ch/en/dokumentation_aviatik_berichte_ueber_unfaelle_schwere_vorfaelle_suchen.html

Baker, D. P., & Dismukes, R. K. (2002). A framework for understanding crew performance assessment issues. *The International Journal of Aviation Psychology, 12*(3), 205-222. http://dx.doi.org/10.1207/S15327108IJAP1203_2

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6*(2), 205-212.

Borman, W. C. (1978, April). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology*, *63*(2), 135-144. http://dx.doi.org/10.1037/0021-9010.63.2.135

Brannick, M. T., Prince, C., & Salas, E. (2002). The reliability of instructor evaluations of crew performance: Good news and not so good news. *The International Journal of Aviation Psychology, 12*(3), 241-261. http://dx.doi.org/10.1207/S15327108IJAP1203_4

Deaton, J. E., Bell, B., Fowlkes, J., Bowers, C. A., Jentsch, F., & Bell, M. A. (2007). Enhancing team training and performance with automated performance assessment tools. *The International Journal of Aviation Psychology, 17*(4), 317-331. http://dx.doi.org/10.1080/10508410701527662

Dekker, S. W. A., Nyce, J. M., van Winsen, R., & Henriqson, E. (2010). Epistemological self-confidence in human factors research. *Journal of Cognitive Engineering and Decision Making, 4*(1), 27-38. http://dx.doi.org/10.1518/155534310X495573

Dismukes, R. K., McDonnell, L. K., & Jobe, K. K. (2009). Facilitating LOFT debriefings: Instructor techniques and crew participation. *The International Journal of Aviation Psychology, 10*(1), 35-57. http://dx.doi.org/10.1207/S15327108IJAP1001_3

Dutra, L., Norman, D., Malone, T., McDougall, W., & Edens, E. (1995). Crew resource management/assessment: Identification of key observable behaviours. In R. Jensen, & L. Rakovan (Eds.), *Proceedings of the 8th Symposium of Aviation Psychology* (pp. 562-567). Columbus, OH: Ohio State University.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (revised ed.). Cambridge, Mass: MIT Press.

Flin, R., & Martin, L. (2001). Behavioral markers for crew resource management: A review of current practice. *The International Journal of Aviation Psychology, 11*(1), 95-118.

Flin, R., Martin, L., Goeters, K.-M., Hörmann, H.-J., Amalberti, R., Valot, C., & Nijhuis, H. (2003). Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills. *Human Factors and Aerospace Safety, 3*(2), 95-117. http://dx.doi.org/10.1207/S15327108IJAP1101_6

Goldsmith, T. E., & Johnson, P. J. (2002). Assessing and improving evaluation of aircrew performance. *The International Journal of Aviation Psychology, 12*(3), 223-240. http://dx.doi.org/10.1207/S15327108IJAP1203_3

Hamman, W. R., & Holt, R. W. (1997). Line operational evaluation (LOE) air carrier scenario based evaluation. *Proceedings of the Human Factors and Ergonomics Society, 41st Annual Meeting* (pp. 907-911). Albuquerque, NM: HFES.

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

Holt, R. W., Hansberger, J. T., & Boehm-Davis, D. A. (2002). Improving rater calibration in aviation: A case study. *The International Journal of Aviation Psychology, 12*(3), 305-330. http://dx.doi.org/10.1207/S15327108IJAP1203_7

Holt, R. W., Johnson, P. J., & Goldsmith, T. E. (1997). Application of psychometrics to the calibration of air carrier evaluators. *Proceedings of the Human Factors and Ergonomics Society, 41st Annual Meeting* (pp. 916-920). Albuquerque, NM: HFES.

Mavin, T. J., & Dall'Alba, G. (2011). Understanding complex assessment: A lesson from aviation. Proceedings of the *4th International Conference of Education, Research and Innovation* (pp. 6563-6570). Madrid, Spain.

Mavin, T. J., & Roth, W.-M. (2014). A holistic view of cockpit performance: An analysis of the assessment discourse of flight examiners. *The International Journal of Aviation Psychology, 24*(3), 210-227. http://dx.doi.org/10.1080/10508414.2014.918434

Mavin, T. J., Roth, W.-M., & Dekker, S. W. A. (2013). Understanding variance in pilot performance ratings: Two studies of flight examiners, captains, and first officers assessing the performance of peers. *Aviation Psychology and Applied Human Factors, 3*(2), 53-62. http://dx.doi.org/10.1027/2192-0923/a000041

Mulqueen, C., Baker, D. P., & Dismukes, R. K. (2002). Pilot instructor rater training: The utility of the multifacet item response theory model. *The International Journal of Aviation Psychology, 12*(3), 287-303. http://dx.doi.org/10.1207/S15327108IJAP1203_6

Norman, G. (2010, December). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15*(5), 625-632. http://dx.doi.org/10.1007/s10459-010-9222-y

O'Connor, P., Hörmann, H.-J., Flin, R., Lodge, M., Goeters, K.-M., & The JARTEL Group (2002). Developing a method for evaluating crew resource management skills: A European perspective. *The International Journal of Aviation Psychology, 12*(3), 263-285. http://dx.doi.org/10.1207/S15327108IJAP1203_5

O'Connor, P., Jones, D. W., McCauley, M. E., & Buttrey, S. E. (2012). An evaluation of the effectiveness of the crew resource management programme in naval aviation. *International Journal of Human Factors and Ergonomics, 1*(1), 21-40. http://dx.doi.org/10.1504/IJHFE.2012.045272

Roth, W.-M., & Mavin, T. J. (2013). Assessment of non-technical skills: From measurement to categorization modeled by fuzzy logic. *Aviation Psychology and Applied Human Factors, 3*(2), 73-82. http://dx.doi.org/10.1027/2192-0923/a000045

Roth, W.-M., Mavin, T. J., & Munro, I. (2014). Good reasons for high variability (low inter-rater reliability) in performance assessment: Toward a fuzzy logic model. *International Journal of Industrial Ergonomics, 44*(5), 685-696. http://dx.doi.org/10.1016/j.ergon.2014.07.004

Salas, E., Bowers, C. A., & Prince, C. (1998). Special issue on simulation and training in aviation. *The International Journal of Aviation Psychology, 8*(3), 195-196. http://dx.doi.org/10.1207/s15327108ijap0803_1

Salas, E., Wilson, K. A., Burke, C. S., & Wightman, D. C. (2006). Does crew resource management training work? An update, an extension, and some critical needs. *Human Factors, 48*(2), 392-412. http://dx.doi.org/10.1518/001872006777724444

Thomas, M. J. W. (2004). Predictors of threat and error management: Identification of core nontechnical skills and implications for training systems design. *The International Journal of Aviation Psychology, 14*(2), 207-231. http://dx.doi.org/10.1207/s15327108ijap1402_6

van Alphen, A., Halfens, R., Hasman, A., & Imbos, T. (1994, July). Likert or Rasch? Nothing is more applicable than good theory. *Journal of Advanced Nursing, 20*(1), 196-201. http://dx.doi.org/10.1046/j.1365-2648.1994.20010196.x

Woehr, D. J., & Huffcutt, A. I. (1994, September). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*(3), 189-205. http://dx.doi.org/10.1111/j.2044-8325.1994.tb00562.x