UNIVERSITY OF GOTHENBURG
DEPARTMENT OF PSYCHOLOGY

The Presence and Impact of Order Effects in Airline Pilot Competency Assessments

Mark Milich

Individual Paper, 15 Credits
Bachelor Thesis in Psychology
PX1500
Spring Term 2019

Supervisor: Lars-Olof Johansson

The Presence Impact of Order Effects in Airline Pilot Competency Assessments

Mark Milich

Abstract. This study investigated whether competency assessments of professional Airline Pilots are subject to order effects such as primacy or recency effects. 18 examiners participated in a web-based experiment where they evaluated crew performance in three short video-recordings of a flight crew performing in a simulator. The scenarios, depicting three levels of competency, were randomly presented in either an improving or deteriorating order. A final, overall evaluation was made following the three scenarios. Results indicate no recency effects, although weak statistical signs of systematic differences in gradings between presentation orders exists. A low number of participants combined with high variations stipulates a cautious conclusion that order effects are likely to have been present and may skew the appraisal of true performance.

In many professions, performance and competency assessments are natural and integrated parts of working life. This is especially true when it comes to professional airline pilots, who are assessed on their competence on a minimum of three occasions annually (European Commision, 2012). Regardless whether the purpose of workplace competency assessments is for safety assurance, career path development, licensing requirements etc, it is nonetheless a reasonable desire that assessments are fair, accurate and reliable. However, assessments carried out by individual assessors are fallible and subject to errors such as observational inaccuracy and cognitive biases (Morgeson & Campion, 1997). Performance evaluations are complex cognitive tasks, consolidating several mental computations. These processes comprise of observing, directing attention, detecting and selecting relevant behaviours, categorizing them in the working memory and then make comparisons towards prevalent professional norms and standards (Tavares & Eva, 2013). As long as the evaluation process takes place, these cognitive operations are performed for a multitude of stimuli simultaneously. However, with increasing complexity follows an increasing risk that assessors revert to mental shortcuts and heuristic principles that may limit or bias the evaluations (Tavares & Eva, 2013).

Even in settings with predefined competency standards such as within medical education and flight crew performance assessments, previous studies recognises that there are indeed considerable variances in the way assessors appraise job performance and competency (Gontar & Hoermann, 2016; Kogan, Conforti, Bernabeo, Iobst, & Holmboe, 2011; Roth, Mavin, & Munro, 2014). There are many factors that explain assessors' varying judgements such as *frame of reference* (Kogan et al., 2011), attention and observational errors (Tavares & Eva, 2013) or pre-occupancy with controlling simulations or training scenarios (Baker & Dismukes, 2002).

Among the variability of cognitive errors that may occur is the tendency amongst assessors to compare with, and become biased by recently observed performance (Yeates, O'Neill, Mann, & Eva, 2012). The sequence in which a number of personality traits are described have proved to have an impact of the overall evaluation of a personality (Asch, 1946). Later studies have dealt with order and sequence effects in job interviews (Farr, 1973), memory research (Greene, 1986) and competency assessments (Murphy, Balzer, Lockhart, & Eisenman, 1985; Yeates et al., 2012). A number of various terms and concepts have been employed to illustrate these effects. In the context of competency and performance assessments these have

been operationalized as for example in the *primacy effect* (Murphy et al., 1985), *recency effect* (Steiner & Rain, 1989), *anchoring* and *contrast bias* (Yeates et al., 2012).

The primacy effect as explained by Asch (1946) implies that an impression that is formed on information acquired early in the judgement process, will be given more weight than information acquired later on. Further studies on primacy effect suggests that the initial stimuli may affect the subsequent information retrieval and thus have an impact on the perception of both the subsequent stimuli and the overall experience (Duffy & Crawford, 2008). Murphy et al (1985) describes the *assimilation effect* and the *contrast effect* as being two possible outcomes of an order effect following an initial stimulus. The assimilation effect biases the judgement towards the previous evaluation and thus evens out a change in performance level. The contrast effect acts in the opposite direction by increasing the assessors' attention to a change in performance, resulting in a higher difference in perceived performance between the two subsequent events. For example, say that an assessor is exposed to two successive stimuli, where the first is negative and the second is of neutral characteristic. If the assimilation effect is present, the assessor will judge the second stimuli as *less* favourable; however, if the contrast effect is present, the judging of the second stimuli will be *more* favourable.

As presented by Murphy et al. (1985), several theories can explain the assimilation effect. One theory suggests that previous impressions are being integrated with present observations. Another theory proposes that confirmation bias or priming causes biases in attention and encoding of observations. A final explanation is that an assessor will, after observing a sequence of consistent performances, continue to award coherent ratings even though the performance has changed. Contrast effects, on the other hand, have been attributed to adaption theory. Here, an assessor having adapted to a specific level of performance and thus reacts to deviations from this level. Yet an explanation is that inconsistent performances are more readily remembered and therefore influences the evaluation to a higher degree. Whether the assimilation effect or contrast effect will be dominating in an assessment situation is generally difficult to predict (Murphy et al., 1985). The cognitive demands stipulated by the actual situation determines which of these two effects that will come in force. In a related study (Murphy, Gannett, Herr, & Chen, 1986), findings suggested that assimilation effects were present when attention was minimized and memory demands where high, while contrast effects were present with high degree of attention and low memory demands.

The tendency to resort to the primacy effect has in previous studies been shown to be associated with high *Need for Cognitive Closure* (NFC), (Freund, Kruglanski, & Shpitzajzen, 1985; Tomić, Tonković, & Ivanec, 2017). NFC has been defined as the desire to come to an answer on a given topic in contrast to remaining in uncertainty (Kruglanski, Webster, & Klem, 1993). "An answer" is here emphasized as being *any* answer. High NFC can either stem from individual traits or be a product of situational factors such as time constraints (Kruglanski et al., 1993; Webster & Kruglanski, 1994)

The theories behind the recency effect mostly deal with research on memory and retrieval (Greene, 1986). The general explanation suggests that the last encountered observation is easier to recall (Costabile & Klein, 2005). Evidence of recency effects were found in performance evaluations of lecturers (Steiner & Rain, 1989) and in juror judgements (Costabile & Klein, 2005). Both studies suggest that the last encountered information will have greater impact on the final decision. Other studies have suggested that recency effect is a function of accessibility (Richter & Kruglanski, 1998). This study by Richter and Kruglanski (1998) also concluded that high NFC could be correlated with recency bias.

Anchoring was described by Tversky and Kahneman (1974) as a cognitive heuristic in which people where influenced too heavily by the first piece of information acquired in a judgement situation. Anchoring bias has been studied in contexts such as medical educational settings (Yeates et al., 2012), clinical assessments (Aronoff, 1999) and in judicial decision

making (Englich, Mussweiler, & Strack, 2006). Research on anchoring effects shows that even an irrelevant or randomly chosen stimulus can influence a succeeding decision (Englich et al., 2006). The mechanism behind anchoring has been described as being one of two possible cognitive processes (Kahneman, 2011). Either an unconscious process where the anchor serves as a priming agent, or a conscious process where the anchor, which is known by the individual to be inaccurate, is subject to a conscious adjustment away from the anchor. In this latter case, the adjustment is generally insufficient, resulting in a judgement too close to the initial anchor. Mussweiler and Strack (1999) have suggested *Selective Accessibility* to account for the anchoring effect. This model suggests a two-step model where anchor consistent information is more rapidly and readily assessed than anchor inconsistent information. An experiment in judicial decision making (Englich et al., 2006) validated this model by measuring the response time of a decision making task where the participants had been primed with either a consistent or a non-consistent anchor. Those decisions which had been preceded by a consistent anchor were on average completed faster. However, this effect was only significant where the given information was of incriminating or negative nature.

In the context of competency assessments, this last finding raises the question whether order effects would differ in situations with weak/negative or good/positive performance. In fact, there are several studies that confirm negativity bias as a condition where negative information receives more attention, quicker identification and have more influence in decision making than positive information (Dijksterhuis & Aarts, 2003; Rozin & Royzman, 2001). This bias has also been established in research on interview and selection processes (Schmitt, 1976). Different levels of recency effect in performance assessments was found depending on whether the performance was weak or good (Steiner & Rain, 1989). These results are though somewhat contradictory as the recency effect after a poor performance was strongest only when the exposure had been distributed over a few days. Where the exposure had been concentrated to just one day, the significant recency effect was found only after observing a good performance.

In summary, we can conclude that even while the psychological mechanisms may be somewhat different, the primacy effect, assimilation effect and anchoring are typically working in the same direction, i.e. skewing a succeeding evaluation to resemble the previous judgement. Contrast bias works in the opposite direction by exaggerating the differences in subsequent performance. When a sequence of varying performances has been observed, the recency effect may have an overall greater impact on a final judgement.

The scope of this paper is to explore the presence and impact of these order effects in the context of airline pilot competency assessments.

Competency Assessment of Commercial Airline Pilots

In terms of competency and suitability, few professionals are as closely examined and assessed as commercial airline pilots (Goldsmith & Johnson, 2002). The primary objective of competency assessments in the airline industry is to ensure a sufficient level of skills and professionalism for the safe transport of the travelling public. Pilots in Commercial Air Transport are assessed on their competence in a flight simulator every 6 months, and during a regular line flight every 12 months, summing up to three competency assessments per year (European Commision, 2012). These evaluations are performed by a single authorized examiner who assesses the flight crew according to a predetermined template (European Aviation Safety Agency, 2019) and the outcome is reported to the national aviation authority as well as the internal airline training department.

While the ultimate purpose of these assessments is to ensure a high flight safety standard, these records may be also be used for varying purposes such as individual career development,

monitoring of training efficiency or to indicate individual training needs (Goldsmith & Johnson, 2002). Many types of grading scales exist ranging from dichotomous pass/fail to multi-grade scales. For the issue or renewal of a license by the licensing authority, only the dichotomous pass/fail-outcome has to be reported (European Aviation Safety Agency, 2016). However, for internal use within each airline the multiple-grade scales may be used at company discretion (International Air Transport Association, 2013). During the writing of this paper, at least four larger European air carriers are known that employs a multi-grade scale. Common for these is that only the lowest grade implies a "fail", with resulting additional training. The remaining steps are various degrees of acceptable performance, however with the implications that only pilots with an overall grade above a specific middle level from recent assessments, can be considered for career movements such as promotion to captain or instructor. Another usage of grading scales is employed in the upcoming transition to the *Evidenced Based Training* model that the aviation industry is currently facing (International Air Transport Association, 2013). Under this new paradigm, one of the changes is that during recurrent training of airline pilots, operators shall track and identify individual pilots training needs. A proposal from European Aviation Safety Agency (2018) suggests a 5-point grading scale. In this proposal the grade 2 is still acceptable, but if this grade is achieved twice for the same competency unit after two following training cycles, i.e. 6 months apart, this should indicate a training need and has to be addressed.

Even though flight examiners generally has a systematic approach to the evaluation process, studies have shown that there is a considerable amount of variation, both in terms of *what* is being observed, how it is interpreted and subsequently evaluated, weighted and finally graded (Roth, 2016). The different performance items that are assessed are usually classified as either Technical Skills or Non-Technical Skills. Technical Skills are skills relating to knowledge, correct application of procedures or checklists, manual flying skills, appropriate use of automation, etc. Non-technical skills refers to areas such as communication, leadership, decision making etc. (Flin et al., 2003). Technical Skills are generally viewed as being more precisely evaluated than Non-Technical Skills (Brannick, Prince, & Salas, 2002; Mavin, Roth, & Dekker, 2013) as these can be described in more concrete terms. However, it is also argued that in the highly complex environment of a modern airliner, the separation between technical and non-technical skills are becoming less distinguishable from each other, affecting each other (Mavin et al., 2013) and thereby eroding the strict separation between technical or non-technical skills as stand-alone performance items. In general, it has been stated that the overall inter-rater reliability shows large variances in pilot competency assessments (Roth et al., 2014).

Several causes for high variation and low interrater reliability in aviation training has been discussed. Some of these can be described as either situational or organisational factors. The simulator environment has been partly attributed to why many examiners differ in which details they have actually observed and taken notice of from a specific session (Baker & Dismukes, 2002). During a simulator scenario instructors and examiners have to manipulate simulator settings while running the simulation and thus cannot devote full attention to observation and evaluation (Baker & Dismukes, 2002). Other factors has been attributed to the grading and evaluation rubrics (Baker & Dismukes, 2002; Goldsmith & Johnson, 2002). Wording, phrasing and the level of abstraction of the grade sheets have shown to play a role in the reliability of aircrew performance assessment (Brannick et al., 2002; Gontar & Hoermann, 2016; Holt, Hansberger, & Boehm-Davis, 2002).

Surprisingly though, little attention has so far been given towards various sources of cognitive aspects of rater reliability in aviation training. A literature research revealed no previous studies that directly examines cognitive biases as causes for misjudgements or high variance in aircrew performance assessments. Rater bias and cognitive effects are indeed known occurrences though not specifically in the research focus. The complications of rater-based

assessments are indicated in several studies. Roth (2016) describes flight examiners interpretations as being "sometimes correct or sometimes more fuzzy" (p. 224) and describes situations where examiners are making early observations leading to a sense, or a narrative, that later on determines which behaviours to attend to. Goldsmith and Johnson (2002) reveals that *halo effects* are a known phenomenon and Roth and Mavin (2013) describes flight examiners work as "inherently imprecise and uncertain" (p. 75). Cognitive biases such as the halo-effect, primacy/recency effects, saliency and confirmation bias has also been pointed out by Arendt (2007) as possible sources for distortion of observed performance. Despite this, there appears to be a paucity of research that directly addresses the cognitive aspects of low interrater reliability.

The current study aims to supplement existing research on inter-rater reliability and rater-bias in aviation training by exploring whether order effects such as primacy and recency effects, anchoring, assimilation and contrast bias may be present during simulator assessments of pilots' proficiency. A simulator evaluation session of today comprises a multitude of mandatory items to be covered (European Aviation Safety Agency, 2016). A crew of two pilots are typically scheduled for a four hour session (Roth, 2016) with high intensity and during which a considerable number of emergency and non-normal events are trained and evaluated. During this period, it may be considered natural that performance varies across the session with temporary ups and downs, caused by brief stress, fatigue or the mental tension of being assessed. The question here asked is whether the sequencing of these variations of high and low performance can affect both the grading of separate performance items as well as the overall impression and grading of a pilot's competency. If this is indeed true, it has the possible implication of distorting and invalidating the fundamental idea behind programmes such as Evidence Based Training and any internal use of pilot assessment results. In the longer term, it may also misdirect training efforts, induce unnecessary monetary training investments or engender unfair treatment of individual pilots.

Two main hypotheses are set up for this study. The first hypothesis, (H1), postulates that: A series of events containing various levels of performance, will, if witnessed and evaluated in two different orders, produce different gradings on each and one of these events, due to order effects. This hypothesis is non-directional, as this difference could either take the form of an assimilation effect, a contrast effect or an anchoring and adjustment effect.

The second hypothesis, (H2), relates to recency effects, and postulates that: The last observed performance level, will have a greater impact on a subsequent overall grading, covering the whole scenario as an entity, distributed at the end of the observations.

Method

Participants

In total 94 participants were invited to take part in this experiment. The participants were all professional flight examiners authorized by the Swedish, Danish and Norwegian Civil Aviation Agencies. The participants were contacted by e-mail through the official contact addresses provided by the relevant aviation authority. All invited participants were licensed Type Rating Examiners (TRE's) qualified to conduct competency assessments on Boeing B737 aircraft. A total number of $N = 19$ participants responded to the full survey. Another 20 examiners visited the survey web-page but did not continue beyond the introductory questions. Otherwise no partial responses were recorded.

The respondents were all men. 11 were aged 56 or older, 7 were between 41 and 55 and 1 was between 31 and 40. 17 had more than 20 years of experience as commercial pilots, 2 had

between 10 and 20 years of experience. 9 examiners had more than 20 years' experience as instructor or examiner, 8 had between 10-20 years and 2 had between 6-10 years.

Instrument

A web-based survey tool was used for this experiment. The surveys consisted of three separate video recordings of a crew of two pilots performing three shorter scenarios. Between each scenario the examiners were asked to rate the two pilots' respective performance according to a 5-point grading scale; grade 1 being the lowest and grade 5 the highest. Grade 1 to 5 where respectively labelled as Poor, Fair, Average, Above Average and Excellent. After the third and final scenario the examiners were asked to rate the pilots' overall performance using the same scale and decide whether the pilots passed or failed the session as a whole.

The video recordings were all filmed in a high-fidelity flight simulator used for advanced flight training. The simulator incorporated a full size, fully equipped cockpit of a Boeing B737 aircraft. Two actors consisting of two licensed and experienced airline pilots qualified on the B737 model were flying the simulator through the scenarios as would be done in a real-life simulator competency assessment. One of the pilots was acting as Captain and the other as First Officer (or co-pilot), both being accordingly uniformed with epaulettes showing their respective rank. All communication was performed in English language as per common industry norm. The pilots were acting according to a predetermined script and were deliberately performing with various proficiencies through the three different scenarios. Each recording was between 2-5 minutes of duration and the flight crew performance were either portrayed as "fair", "neutral" or "good" in the three different clips. Total duration of the recordings were 11 minutes and 50 seconds.

The scenario where the crew would display a "fair" performance started out with a take-off in which the crew experienced an engine "severe damage" failure. The crew subsequently had to perform the associated emergency drill whilst airborne. During this scenario the crew intentionally represented a few typical errors that a real flight crew could easily make. In order not to polarize the performance to the extremes, most of these errors and mistakes were eventually corrected, however not in an ideal manner.

In the "good" performance scenario, the crew were experiencing a complete electrical failure which suddenly disabled several systems and instruments. This is a situation which requires accurate and prompt action as several systems such as autopilot and flight instruments on the First Officers panel immediately fails. For this scenario the emergency drills were accomplished exemplary. Once again, to avoid being overly obvious and to avoid a roof effect, the performance was nuanced to include some minor but negligible lapses.

In between the good and the fair performance scenarios, an excerpt of a "neutral" event was inserted. In order to produce a stimulus that would be considered as just "standard" or "average" performance, a scenario with relatively few challenges was depicted. This way, it would not be possible to expect neither an explicitly good nor poor performance from the crew as long as they simply executed the manoeuvre according to standard procedures. The selected scenario included an approach to landing which had to be aborted. The "Go-Around" manoeuvre that was completed is a standard manoeuvre that by most professional pilots should be executed according to a well-rehearsed procedure.

In the survey, the participants were randomised to view the three scenarios in two different presentation orders. Either the presentation order was of improving performance level, with scenarios being presented in the order of Fair-Neutral-Good, hereby named presentation order FNG (n = 9), or with a reversed, decreasing performance level of Good-Neutral-Fair, called presentation order GNF (n = 10).

Between each transition, the participants were asked to rate both the Captain's and First Officer's performance on three different competency categories plus an overall impression. The first category was labelled "Skills, Knowledge and Procedures", resembling all technical aspects of the aircraft operation, also known as "Technical Skills". The second category was labelled as "Situation Awareness"; a well-established performance dimension within aviation training, referring to the pilots awareness of factors such as: time, position (including speed, altitude and energy state), aircraft system configuration, external environment, anticipation of future events etc, (Flin et al., 2003). The third category was labelled "Crew Resource Management", (CRM), referring to "Non-Technical Skills" (i.e. social and cognitive skills) required for the safe operation of an aircraft, comprising of elements such as co-operation, communication, leadership/followership and decision making (Flin et al., 2003). Once the final scene had been viewed and rated, the participants were asked to give an overall grading to each of the two crew members, followed by an overall pass/fail decision. Accordingly, 13 gradings per each crewmember were distributed by each examiner; four performance dimensions in each of the three scenarios plus the global evaluation. This equals up to a grand total of 26 different evaluations. This, seemingly complex assessment process, was in fact designed to reassemble the natural setting as closely as possible. A grading matrix as used in the survey is presented in Appendix 1.

A final stage of the survey, incorporated four questions relating to Need for Cognitive Closure. In general, a larger set of questions would be required to get an accurate estimate of these dimensions. However, to avoid survey fatigue, the number of questions had to be limited. The four questions were variations of question no's 2, 16, 17 and 21 from the 42-item NFCS scale developed to assess individual differences in Need for Cognitive Closure, (for an overview, see Kruglanski et al., 1993). The wording of the original NFCS questions had been altered slightly to refer to the competency assessment process specifically, e.g. an original wording of *"I usually make important decisions quickly and confidently"* was altered to *"I usually make grading decisions quickly and confidently"*. The questions were answered on a 6-point Likert scale where respondents chose between 1 = *strongly disagree*, 2 = *moderately disagree*, 3 = *disagree,* 4 = *agree*, 5 = *moderately agree* and 6 = *strongly agree*. The NFCS questions as used in this survey are presented in Appendix 2.

No hypothesis was formed around the NFC variable. The reason for the NFCS scale to be included in the survey was to be able to control for an additional confounding variable which is known to have influence on order effects. As it turned out, the response rate was low and therefore this opportunity was not utilized other than to test for a random difference between the two groups FNG and GNF.

Procedure

The participating examiners were invited to take part in the survey by an e-mail function in the survey program. The e-mail informed them on the general purpose of the survey but participants were kept naïve on the specific details. On the start page of the survey participants were informed that the survey was best suited for computers or tablets rather than mobile phones. This was in order to prevent participants from completing the survey without being able to pay full attention to the crew performance as well as being able to observe details that was deemed too subtle to be noticed on a smaller screen. A majority of the invited participants that logged on to the survey never continued beyond this page.

After some introductory questions regarding experience level, age and gender, yet another "gatekeeper" page was inserted just before the video scenarios. This page instructed the participants that the remaining part should be completed in one session without possibilities

to step backwards. All participants who completed the survey did so in one sitting once continuing past this page.

The participants were then shown the three recorded simulator scenarios, which were randomised to any of the two presentation orders FNG or GNF. Just before each scenario started the participants could read a brief outline of the specific topic selected for the upcoming scenario in order to be prepared for which details to pay attention to, e.g "*The crew will shortly experience a Loss of Both Engine Driven Generators which is resettable on one side*". Once the scenario was completed the examiners could take their time to distribute the gradings according to the grading matrix.

Average response duration was $M = 23.5$, $SD = 8.7$ minutes. This figure excluding three responses that were incorrectly timed as being excessively long as the survey timer continued even while the respondents were pausing at the "gatekeeper" page. All responses were assured to be longer in duration than the total duration of the video recordings (11 minutes and 50 seconds).

Response rate were slow and after one week a reminder was sent to unfinished respondents. The survey was open for two weeks before the final data collection began.

Results

The central thesis in this study is whether the sequence in which an aviation examiner observes a series of varying high or low performance may have an impact on the grading of subsequent performance items as well as a final, overall grading. The three recordings were manipulated in order to represent three levels of crew performance; fair, neutral and good. In each scenario the examiners evaluated four different dimensions of individual pilots' performance; Technical Skills, Situation Awareness, CRM and an Overall Impression. A final, global evaluation, covering the three scenarios as a whole, was also made once all three scenarios had been evaluated individually.

The results indicate large variations between examiners' evaluations in each of these four performance dimensions. Gradings in the Fair scenario ranges between grade 1 and 4 with mean values of single performance dimensions between $M_{min} = 2.2$ to $M_{max} = 2.8$, $SD_{min} = 0.7$, $SD_{max} = 1.0$. Gradings in the Neutral scenario also ranges between grade 1 and 4 with mean values ranging from $M_{min} = 2.8$ to $M_{max} = 3.2$, $SD_{min} = 0.7$, $SD_{max} = 1.0$. The Good scenario indicates the highest degree of agreement with grades ranging from grade 3 to 5 and mean values from $M_{min} = 3.6$ to $M_{max} = 4.0$, $SD_{min} = 0.5$, $SD_{max} = 0.8$.

Cronbach's Alpha of the four performance dimensions were between $\alpha = .88$ to .96 across all three scenarios for both the Captain and First Officer. When filming the scenarios, the attempt was to keep these performance dimension congruent with each other and thus a high value was expected. Consequently, in all subsequent analyses these four performance dimensions are averaged.

Average gradings for both crewmembers combined and regardless of presentation order are $M = 2.5$, $SD = 0.63$ for the Fair scenario, $M = 3.0$, $SD = 0.73$ for the Neutral scenario and $M = 3.7$, $SD = 0.56$ for the Good scenario. A manipulation check in order to establish whether these three scenarios can be viewed as inherently different from each other was performed by an ANOVA with repeated measures. The results indicates a significant main effect by scenario ($F(2,36) = 26.71$, $p < .001$, $\eta^2_p = .60$). A Bonferroni corrected post hoc test at $p < 0.05$ revealed that a difference exists between both the Fair and Neutral scenarios ($p = .010$, $d = 0.71$, 95% CI [0.05, 1.36]), as well as between Neutral and Good ($p < .001$, $d = 1.12$, 95% CI [0.43, 1.80]), and between Fair and Good ($p < .001$, $d = 2.02$, 95% CI [1.24, 2.80].

Primacy effects. To identify the presence of primacy effects or associated effects such as assimilation, contrast or anchoring effects, it was hypothesized that there would be a systematic difference between how a specific scenario was evaluated depending on its' preceding stimulus. Either a contrast effect, exaggerating a change in performance, or an assimilation effect, underestimating the change in performance, was expected to occur. The average grades for each crewmember and scenario are displayed in Table 1.

Table 1

Mean gradings for Captain and First Officer respectively in each scenario Fair, Neutral and Good. "FNG" and "GNF" represents presentation order "Fair – Neutral – Good" or "Good – Neutral – Fair".

	Presentation order			
	FNG (n = 9)		GNF (n = 10)	
Scenario/Rank	M	(SD)	M	(SD)
Captain				
Fair	2.6	(0.9)	2.3	(0.5)
Neutral	2.9	(0.6)	3.3	(1.0)
Good	3.5	(0.6)	4.1	(0.6)
First Officer				
Fair	2.9	(0.6)	2.4	(0.7)
Neutral	2.7	(0.6)	3.0	(0.8)
Good	3.5	(0.6)	3.8	(0.6)

The data shows that the average grading for the Fair scenario is lower in the GNF-condition, i.e. when this scenario is observed last in a sequence following a higher performance. Similarly, the average gradings for the Good scenario are also lower when observed last after a series of weaker performance, as in the FNG-condition. In the neutral scenario the averages differ slightly; however, with slightly greater mean standard deviations than in the other scenarios.

To test for a systematic difference in this trend, a 2 (Presentation order: FNG vs. GNF) by 2 (Rank: Captain vs. First Officer) by 3 (Scenario: Fair vs. Neutral vs. Good) mixed ANOVA with repeated measures on the two last factors was carried out in SPSS. In this analyze, the Rank variable has been included as a manipulation check in order to test whether the two crewmembers are evaluated significantly different from each other as the intent in the scenarios was to keep the two pilots' performance levels congruent with each other.

The results revealed a significant main within-subjects effect of Scenario ($F(2,34) = 31.22$, $p < .001$, $\eta^2_p = .65$) as well as a significant interaction effect of Presentation order by Scenario ($F(2,34) = 4.86$, $p < .05$, $\eta^2_p = .22$). No main or interactions effects were detected involving rank and thus data covering both crew members can continue to be analyzed jointly.

A Bonferroni corrected post hoc paired samples t-test at $p < 0.05$ revealed that significant differences were found between scenario N ($M = 3.2$, $SD = 0.8$) and G ($M = 3.9$, $SD = 0.5$, $p = .006$, $d = 1.14$, 95% CI [0.20, 2.09]), and between scenario F ($M = 2.3$, $SD = 0.5$) and G ($M = 3.9$, $SD = 0.5$, $p < .001$, $d = 3.12$, 95% CI [1.82, 4.43]) in presentation order GNF and both pilots combined. Table 2 and Table 3 illustrates all nine combinations of post hoc tests.

Table 2

Independent t-tests between presentation order FNG and GNF reported for Captains' and First Officers' Fair, Neutral and Good gradings combined.

	Presentation order							
	FNG (n = 9)		GNF (n = 10)					
Scenario	*M*	(*SD*)	*M*	(*SD*)	*t*(17)	*p*	*d*	95% CI
Crew combined								
Fair	2.7	(0.7)	2.3	(0.5)	1.51	.15	0.69	-0.23, 1.62
Neutral	2.8	(0.6)	3.2	(0.8)	0.99	.34	0.45	-0.46, 1.37
Good	3.5	(0.6)	3.9	(0.5)	1.81	.09	0.83	-0.11, 1.77

Table 2 discloses that despite the significant interaction effect by presentation order, as visualized in Figure 1, data from the pairwise comparison does not reach a significant statistical difference between any of the two presentation orders for any scenario. The Good scenario is the most significant with p = .09.

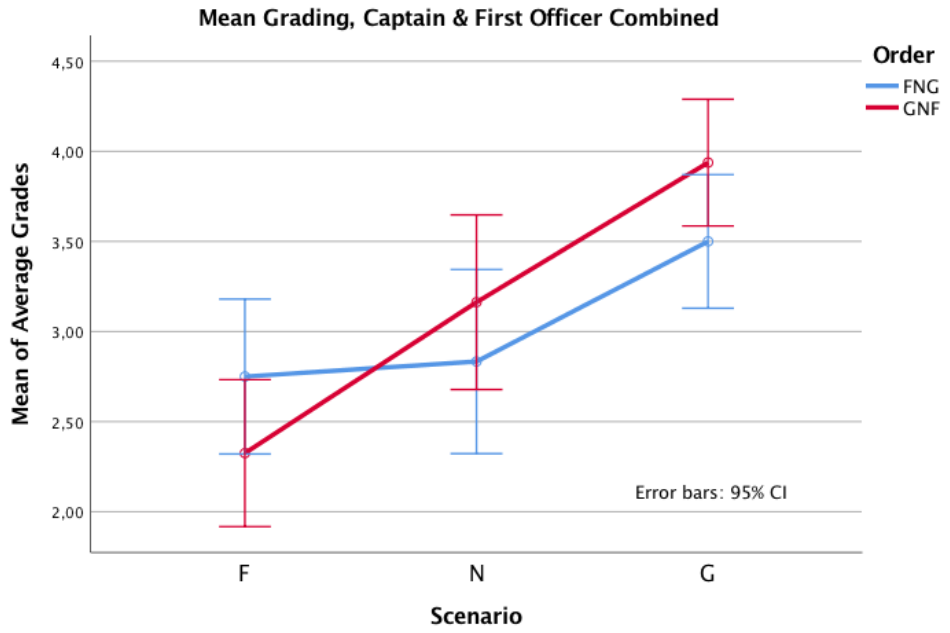

Figure 1: Captains and First Officers combined average grading in each scenario. The figure illustrates main effects by scenario as well as an interaction effect by presentation order. Note that the presentation order GNF is reversed and should be read from right to left for correct sequencing.

Likewise, as seen in Table 3, when dividing the evaluation into the FNG and GNF groups, the significance of the main effect by scenario decreases, especially for the comparison between scenario Fair and scenario Neutral in the FNG presentation order. At the Bonferroni corrected significance level only the GNF presentation order reaches significance, which is limited to the comparisons between Neutral vs. Good and Fair vs. Good.

Table 3

Paired samples t-tests between presentation order FNG and GNF reported for Captains' and First Officers' Fair, Neutral and Good gradings combined. Means and Standard deviations as found in Table 2.

Scenario compared	*t*	*p*	*d*	95% CI
FNG (*n* = 9)				
Fair vs. Neutral	0.60	.567	0.13	-0.79, 1.05
Neutral vs. Good	3.23	.012*	1.15	0.16, 2.15
Fair vs. Good	3.46	.009*	1.19	0.19, 2.20
GNF (*n* = 10)				
Fair vs. Neutral	3.38	.008*	1.20	0.25. 2,15
Neutral vs. Good	3.62	.006**	1.14	0.20, 2.09
Fair vs. Good	7.25	.000**	3.12	1.82, 4.43

** Indicates statistical significance at p < .05.*

*** Indicates Bonferroni corrected statistical significance at p < 0.05/9 = 0.006*

Recency Effect. To test for the recency effect, it was hypothesized that there will be a difference in the global evaluation that were to be distributed just after grading the last scenario and that this grading would be more affected by the last observed performance than the preceding performances. Global gradings for each crewmember is presented in Table 4.

Table 4

Global evaluation: Average overall grade for Captain and First Officer respectively for each presentation order FNG and GNF.

	Presentation order							
	FNG (*n* = 9)		GNF (*n* = 10)					
Rank	*M*	(*SD*)	*M*	(*SD*)	*t* (17)	*p*	*d*	95% CI
Captain	2.9	(0.9)	2.9	(0.6)	0.03	.975	-0.01	-0.90, 0.90
First Officer	2.7	(0.5)	2.7	(0.7)	0.12	.905	-0.05	-0.95, 0.85

An independent samples t-test revealed that no significant difference was found in the global gradings between the two different presentation orders; however, once again large variances between the examiners were evident. A final question in which examiners were asked to either pass or fail the crewmember shows that five out of the 18 examiners failed either one or both crewmembers. The Captain was judged to have failed in four occasions and the First Officer was judged to have failed in two occasions. One of the examiners judged both crew members to have failed simultaneously.

Need for Cognitive Closure. To test the participants for NFC, four modified questions from the NFCS scale (Kruglanski et al., 1993) was used. To check whether there was any difference between the examiners in group FNG (*n* = 9) and GNF (*n* = 10) an independent t-test was performed on NFCS score. The test revealed no significant differences in NFC between the groups. Both groups scored *M* = 15.7, *SD* = 2.0, with subsequent *t*(17) = 0.085, *p* = .933. Maximum possible score on these questions were 24 points. Given the low number of participants no further analyses were made on the NFC results.

Discussion

The purpose of this study was to test whether authorized flight examiners are influenced by order effects such as the primacy and recency effects while conducting competency assessments of professional airline pilots. Two main hypotheses were set up for this study. The first hypothesis concerns the presence and impact of primacy effects during an ongoing evaluation and the second hypothesis concerns how recency effects may impact a summarising global evaluation once the test has been completed.

Primacy effects. For the primacy, assimilation or contrast effects it was hypothesised that the sequence in which a flight examiner observes a series of varying crew performance, will have an impact on all subsequent evaluations. This hypothesis was tested by altering the presentation order of three video recorded simulator scenarios of varying performance. The collected data revealed that there appear to be a trend in systematic difference in how examiners evaluate performance. However, the statistical analyses do not fully support a univocal interpretation as a significant statistical result is not achieved in all necessary tests. Due to the low significance level combined with a low sample size these findings should therefore be interpreted with caution.

Even though the statistical significance may be somewhat weak, the data from this particular study suggests that flight examiners do differ in their evaluations depending on whether the observed performance is of improving or decreasing nature. With reference to the diagram in Figure 1, examiners appear to adapt more rapidly to decreasing performance and more slowly to improving performance. As the diagram indicates, initial evaluation in the decreasing presentation order GNF, starts out with high grades, the second evaluation falls considerably, but are still in average higher than in the FNG condition. In the final evaluation, the Fair scenario, gradings are notably lower than in the FNG condition. A similar, but reversed effect is shown in presentation order FNG. Here, the data indicate that increasing performance has less impact on examiners grading and that sensitivity for increasing performance is lower than for decreasing performance.

There may be several psychological mechanisms playing part in this discrepancy between presentation orders. In the improving presentation order FNG, it can be argued that an assimilation effect is present between scenario F and N, as examiners appears to be biased by the low performance in the first observed scenario. It could also be argued that differences are due to an anchoring and adjustment heuristics. With reference to the diagram in Figure 1; the subsequent evaluations in scenario N and G for both presentation orders, generally discloses a change of performance, but with insufficient adjustments from previous evaluations. With deceasing performance, as in presentation order GNF, it can be argued that we see a contrast effect as examiners are adapting more rapidly to changes in performance, especially in between scenario N and F. The hypothesis in this study postulated that we would see either an assimilation effect *or* a contrast effect. It was therefore rather surprising to find evidence of both, depending on presentation order. An alternative explanation in lieu of assimilation and contrast theories could therefore be that the difference in sensitivity to improving or decreasing performance is in fact due to negativity bias. If examiners are being more attentive to negative performances than positive performances it could also be expected that they will be more sensitive for decreased performance than improved performance.

It should be noted though, that for both presentation orders, variances in the second observed scenario (scenario N), are high. This data could therefore actually contain a random occurrence of both assimilation effects and contrast effects. If this is the case, these effects would be obscured by the large variations found in the data.

To reiterate, the findings are subject to low level of statistical significance, hence the discussion above are of a speculative nature. A last explanation for the differing performance

evaluations could be that examiners while watching the initial scenario, has not yet directed their full attention to the ongoing activity and hence are not fully attentive to details. This could very well be the situation in the presentation order FNG, where, in the Fair scenario, evidence is found of a central tendency bias, i.e. the tendency for a rater to distribute average level grades (Arendt, 2007). However, if this was the case, it would be unclear whether an equal central tendency bias where to be found in the Good scenario in presentation order GNF, instead of the rather high gradings that were now found.

Recency effects. For the recency effect it was hypothesized that the overall grade assessed at the end of the three scenarios would be influenced by the performance in the last observed scenario. The test results indicate no significant difference in global grading between the two presentation orders and thereby no sign of a measurable recency effect. However, in the absence of a recency effect it is more notable that the variance of a global evaluation is almost as large as in each single point performance item. It should be noted though that the total length of the three scenarios only summed up to just about 12 minutes. Whether this short duration would be able to produce a recency effect is questionable. It is likely that this duration is actually brief enough to still be retained in the memory more or less in its' entirety and thereby being equally influenced by both the first and the last observed scenario. As there generally isn't time available in on-line surveys to conduct experiments with longer durations, a solution to this problem in future research could perhaps have been to add a filler task in between the scenarios and the final evaluation. This could be in the form of some of the more general questions asked in the survey.

Another aspect is that while the initial performance items consisted of twelve separate evaluations per pilot, the global grading only consisted of one single evaluation. Furthermore, the grading scale used in the experiment was a 5-point scale, which for this purpose is probably a bit too course and insensitive. However, this model was chosen in order to be consistent with common practice in aviation training, where global evaluations are commonly graded on one single, summarizing dimension using the same scale as in separate items. The combination of a single measurement point and a relatively insensitive scale possibly makes the instrument slightly too imprecise to measure this effect. In hindsight, some more measurement points might have been added to increase sensitivity on this particular dimension.

Additional Findings

One of the most evident findings when looking at the data acquired in this survey is that examiners differ widely in their assessments. In total the examiners distributed 24 different single point gradings plus two global assessments. In nine of these 26 assessments, the total range of gradings varies with four units, one varies with five units and the remaining 14 varies with three units. i.e. a normal distribution covers a considerable part of the grading scale. Only in the Good scenario the examiners appear to be in reasonable agreement with no grades below Average. In all other scenarios the gradings varies between Below Average and Above Average and beyond. This finding is consistent with previous research of Gontar and Hoermann (2016) which also indicated higher agreement for outstanding or high performance. Another finding that complicates and distorts data in this study is that several individual responses shows a lack of sensitivity for changing performance levels.

In a statistical sense this variation might not be of concern. However, in an industry, where competency assessments are a fundamental structural element and as these assessments are determining career development and establishing individual training needs, this should raise concern. It should be pointed out though, that the participants in this experiment are of a random combination from an unknown number of different airlines and training organisations. Each of

these have their own grading scales, standards and norms to which they evaluate their pilots. Even if the scales and vocabular used in this survey to the highest possible degree was of industry norms, slight variations of these are possible and this impact cannot be fully estimated. Nevertheless, this imprecision has been conclusively proved in previous experiments and points to a constitutional susceptibility within aviation training practices.

Limitations

There are several limitations in the present study that needs to be acknowledged. As previously mentioned, the total duration of the three video recordings reaches just about twelve minutes. This should be compared to the natural settings in which a simulator assessment session is typically in the region of 4 hours, preceded by a pre-flight briefing during which the crew may have an oral questioning and perhaps also complete a flight-planning task – all part of the examination. Without further research, it is difficult to predict whether this, significantly longer duration would either reinforce or weaken any primacy- or recency effects. Furthermore, the natural environment is richer on details, contains face to face communication and brings significantly more stimuli to base judgements on. This, on the other hand, also makes the situation more complex, which in turn increases the cognitive demands and possibly induce sources for other biases. In this experiment, examiners only needed to focus on crew performance, while in the natural environment, they would also have to manage simulator settings, control and direct the training scenario, act as air traffic controller, etc. These circumstances though, are similar to those in which other, comparable studies of inter-rater reliability among flight instructors and examiners have been performed.

Yet another possible source of data distortion in this study is the fact that all participating examiners did so completely voluntary on their own spare time and through a web based survey tool. Generally, it could be considered that this condition is not ideal in order to have a complete randomised sample as there could be an inherent difference between examiners who are willing to voluntary participate in web based activities compared to those who don't. In the invitation e-mail it was stated that the study was focusing on inter-rater reliability. Hence, it is possible that those examiners who chose to participate in the experiment was therefore more interested in this topic and therefore perhaps more careful in their evaluations and less susceptible to biases in general. On the other hand, using a web based survey could contrary create a sense that the evaluation "is not for real" and therefore inhibit systematic unbiased reasoning as there are no consequences for committing errors. Other unknown circumstances when considering the validity of this study is the uncontrolled test environment in which the participants have been completing the survey as well as the uncertainty of the participants attention and seriousness with which they conducted the survey.

These are all factors which impact cannot be fully predicted. However, the main weakness in this study lies the low sample size. Given that the participants had to be randomised into two groups, and as the variation within the evaluations are high, the resulting statistical significance therefore turns rather low. Most of the statistical tests applied in this study show non-significant differences between the two presentation orders FNG and GNF; nevertheless, they do point to a systematic difference between the two presentation orders. However, large variances in combination with a small sample of participants makes it perilous to draw conclusions. Findings in this study should therefore be treated with caution.

Conclusion, implications and suggestions for further research

The main scope of this paper was to establish whether order effects such as primacy and recency effects, assimilation and contrast effects and anchoring and adjustment heuristics are present when flight examiners evaluates professional airline pilots. While the statistical significance in this study is generally low, the results in fact do point towards possible evidence of order effects being present. To further establish whether the effects seen here was of a random nature or indeed signs of systematic bias, the study needs to be extended and more responses collected. A recency effect could not be established from this experiment; however, a review of the method used raised questions whether an alternative method would be needed.

Previous findings indicating high variability in crew performance assessments were confirmed in this study. Data in the current research also confirms previous findings that examiners tend to be in higher agreement on good performance, but as performance varies, ambiguity increases and thus also the variation of gradings.

Should the results herein be representative for the industry as a whole, the overall impact of these findings may have some consequential implications. Proposals for enhanced competency assessment programs such as the Evidence Based Training philosophy assumes high rater integrity. European Aviation Safety Agency (2018) stipulates rater standardisation training to be implemented by training organisations although research indicates that rater training are of limited efficiency (Holt et al., 2002; Woehr & Huffcutt, 1994). This raises several questions: Are the impact of cognitive biases among examiners overlooked? Can cognitive biases be addressed in rater training? Is the airline industry overconfident in their belief in competency assessment efficiency and accuracy?

The current study is one of a few studies dealing with rater cognition within the airline training sector. To draw statistical conclusions from this study will require both an extended research method as well as a wider sample size. A general suggestion for further research could be to continue the research in rater cognition and further establish more precisely which cognitive processes that are in force when examiners evaluate flight crew performance. Given the disappointing results in terms of rater reliability and agreement, it would also be welcome with research establishing the organisational and individual effects of low reliability in assessment practices.

References

Arendt, D. N. (2007). *Psychometric properties of aviation safety performance evaluation instruments: Dependability of assessments* (Order No. AAI3228658). Available from PsycINFO. (622007208; 2007-99004-291).

Aronoff, D. N. (1999). *Errors in clinical judgment: The effect of temporal order of client information on anchoring, adjustment, and adjustment mitigation and category of clinical inferences* (Order No. AAMNQ29876). Available from PsycINFO. (619439355; 1999-95003-002).

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41*(3), 258-290. doi:10.1037/h0055756.

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. doi:10.1207/s15327108ijap1203_2

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. doi:10.1207/s15327108ijap1203_4

Costabile, K. A., & Klein, S. B. (2005). Finishing strong: Recency effects in juror judgments. *Basic and Applied Social Psychology, 27*(1), 47-58. doi:10.1207/s15324834basp2701_5.

Dijksterhuis, A., & Aarts, H. (2003). On wildebeests and humans: The preferential detection of negative stimuli. *Psychological Science, 14*(1), 14-18. doi:10.1111/1467-9280.t01-1-01412.

Duffy, S., & Crawford, L. E. (2008). Primacy or recency effects in forming inductive categories. *Memory & Cognition, 36*(3), 567-577. doi:10.3758/mc.36.3.567

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin, 32*(2), 188-200. doi:10.1177/0146167205282152.

European Aviation Safety Agency. (2016). *EASA Part FCL*. Retrieved April 3, 2019, from https://www.easa.europa.eu/sites/default/files/dfu/Part-FCL.pdf.

European Aviation Safety Agency. (2018). *Notice of Proposed Amendment 2018-07(B)*. Retrieved April 11, 2019, from https://www.easa.europa.eu/sites/default/files/dfu/NPA%202018-07%28B%29.pdf.

European Aviation Safety Agency. (2019). *Acceptable Means of Compliance (AMC) and Guidance Material (GM) to Annex III Organisation requirements for air operations [Part-ORO] of Commission Regulation (EU) 965/2012 on air operations*. (2019). Retrieved April 17, 2019, from https://www.easa.europa.eu/sites/default/files/dfu/Consolidated%20AMC-GM_Annex%20III%20Part-ORO_March%202019.pdf.

European Commision. (2012). Commission Regulation (EU) No 965/2012. (2012, October 5). Retrieved April 3, 2019, from https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:296:0001:0148:EN:PDF.

Farr, J. L. (1973). Response requirements and primacy-recency effects in a simulated selection interview. *Journal of Applied Psychology, 57*(3), 228-232. doi:10.1037/h0034708.

Flin, R., Martin, L., Goeters, K., Hörmann, H., 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), 97-119. Retrieved from https://search-proquest-com.ezproxy.ub.gu.se/docview/620169292?accountid=11162.

Freund, T., Kruglanski, A. W., & Shpitzajzen, A. (1985). The freezing and unfreezing of impressional primacy: Effects of the need for structure and the fear of invalidity. *Personality and Social Psychology Bulletin, 11*(4), 479-487. doi:10.1177/0146167285114013.

Goldsmith, T. E., & Johnson, P. J. (2002). Assessing and Improving Evaluation of Aircrew Performance. *The International Journal of Aviation Psychology, 12*(3), 223-240. doi:10.1207/s15327108ijap1203_3

Gontar, P., & Hoermann, H.-J. (2016). Interrater Reliability at the Top End: Measures of Pilots' Nontechnical Performance. *The International Journal of Aviation Psychology, 25*(3-4), 171-190. doi:10.1080/10508414.2015.1162636

Greene, R. L. (1986). Sources of recency effects in free recall. *Psychological Bulletin, 99*(2), 221-228. doi:10.1037/0033-2909.99.2.221.

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. doi:10.1207/s15327108ijap1203_7

International Air Transport Association. (2013). *Evidence-Based Training Implementation Guide, 1st Edition. (2013, March)*. Retrieved from https://www.iata.org/whatwedo/ops-infra/training-licensing/Documents/ebt-implementation-guide.pdf.

Kahneman, D. (2011). Thinking, fast and slow Farrar, Straus and Giroux, New York, NY.

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ, 45*(10), 1048-1060. doi:10.1111/j.1365-2923.2011.04025.x

Kruglanski, A. W., Webster, D. M., & Klem, A. (1993). Motivated resistance and openness to persuasion in the presence or absence of prior information. *Journal of Personality and Social Psychology, 65*(5), 861-876. doi:10.1037/0022-3514.65.5.861.

Mavin, T. J., Roth, W., & Dekker, S. (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. doi:10.1027/2192-0923/a000041.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 82*(5), 627-655. doi:10.1037/0021-9010.82.5.627.

Murphy, K. R., Balzer, W. K., Lockhart, M. C., & Eisenman, E. J. (1985). Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology, 70*(1), 72-84. doi:10.1037/0021-9010.70.1.72.

Murphy, K. R., Gannett, B. A., Herr, B. M., & Chen, J. A. (1986). Effects of subsequent performance on evaluations of previous performance. *Journal of Applied Psychology, 71*(3), 427-431. doi:10.1037/0021-9010.71.3.427.

Mussweiler, T., & Strack, F. (1999). Comparing Is Believing: A Selective Accessibility Model of Judgmental Anchoring, *European Review of Social Psychology, 10*(1), 135-167. doi:10.1080/14792779943000044.

Richter, L., & Kruglanski, A. W. (1998). Seizing on the latest: Motivationally driven recency effects in impression formation. *Journal of Experimental Social Psychology, 34*(4), 313-329. doi:10.1006/jesp.1998.1354.

Roth, W.-M. (2016). Flight Examiners' Methods of Ascertaining Pilot Proficiency. *The International Journal of Aviation Psychology, 25*(3-4), 209-226. doi:10.1080/10508414.2015.1162642

Roth, W.-M., & Mavin, T. J. (2013). Assessment of Nontechnical Skills. *Aviation Psychology and Applied Human Factors, 3*(2), 73-82. doi: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. doi:10.1016/j.ergon.2014.07.004

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296-320. doi:10.1207/S15327957PSPR0504_2.

Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology, 29*(1), 79-101. doi:10.1111/j.1744-6570.1976.tb00404.x.

Steiner, D. D., & Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. *Journal of Applied Psychology, 74*(1), 136-142. doi:10.1037/0021-9010.74.1.136.

Tavares, W., & Eva, K. W. (2013). Exploring the impact of mental workload on rater-based assessments. *Adv Health Sci Educ Theory Pract, 18*(2), 291-303. doi:10.1007/s10459-012-9370-3

Tomić, I., Tonković, M., & Ivanec, D. (2017). Effects of psychological distance and need for cognitive closure on impression formation. *Journal of General Psychology, 144*(1), 1-15. doi:10.1080/00221309.2016.1258385.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. doi:10.1126/science.185.4157.1124.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67*(6), 1049-1062. doi:10.1037/0022-3514.67.6.1049.

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

Yeates, P., O'Neill, P., Mann, K., & Eva, K. W. (2012). Effect of exposure to good vs poor medical trainee performance on attending physician ratings of subsequent performances. *JAMA: Journal of the American Medical Association, 308*(21), 2226-2232. doi:10.1001/jama.2012.36515.

Appendix 1

a) Grading matrix as used after each scenario:

Please grade the crew performance as observed in the scenario

1 - Poor, 2 - Fair, 3 - Average, 4 - Above average, 5 - Excellent

	Captain					First Officer				
	1	2	3	4	5	1	2	3	4	5
Skills, Knowledge & Procedures	○	○	○	○	○	○	○	○	○	○
Situation Awareness	○	○	○	○	○	○	○	○	○	○
CRM	○	○	○	○	○	○	○	○	○	○
Overall Impression (Scenario 1)	○	○	○	○	○	○	○	○	○	○

b) Grading matrix as used for the final Global Evaluation:

You have now observed three short scenarios. Please give your **overall** grade to the crew.

	Captain					First Officer				
	Poor	Fair	Average	Above Average	Excellent	Poor	Fair	Average	Above Average	Excellent
Overall Competency	○	○	○	○	○	○	○	○	○	○

Please indicate whether the crew failed or passed on the session as a whole

	Captain		First Officer	
	Fail	Pass	Fail	Pass
Fail/Pass overall	○	○	○	○

Appendix 2

NFCS questions and associated Likert scale:

- I usually make grading decision quickly and confidently
- I tend to put off deciding which grading to give until the last possible moment
- In most assessments, I can immediately differ between good and poor performance
- When I've made up my mind about a pilots competency, I generally stick to that decision

1.....strongly disagree
2....moderately disagree
3.....slightly disagree
4.....slightly agree
5.....moderately agree
6.....strongly agree