Cultural Practices and Cognition in Debriefing: The Case of Aviation

Wolff-Michael Roth[1,2]

[1]University of Victoria, [2]Griffith Institute for Educational Research

Corresponding author: Wolff-Michael Roth

Email: mroth@uvic.ca

Wolff-Michael Roth is Lansdowne Professor of Applied Cognitive Science at the University of Victoria. He investigates cognition across the life span in formal and informal educational settings and in technology-rich workplaces. His most recent works include *Concrete Human Psychology* (Routledge, 2015) and *Rigor in Qualitative Data Analysis* (Sense Publishers, 2015).

Cultural Practices and Cognition in Debriefing: The Case of Aviation

Abstract This study was designed to investigate the cultural and cognitive dimensions of debriefing. Pilots and flight examiners from five airlines were involved in this cognitive anthropological study. The data include (a) videotaped debriefing sessions and associated interviews with participants and (b) stimulated recall, and modified think-aloud protocols with flight examiners. Findings point to the varied structures and contents of debriefing practices, in part mediated by the tools available to the participants. Some implications from this research are already taken up in the industry and are currently subject to an ongoing investigation.

Keywords debriefing; practices; culture; cognition; embodiment; narrative; gestures

INTRODUCTION

Debriefing is a cultural practice used to reflect on and review, after some action has occurred, preceding events to improve cognition and performance in numerous areas but especially in military (Morrison & Meliza, 1999), medical (e.g., Zuckerman et al., 2012), psychological (Adler, Bliese, McGurk, Hoge, & Castro, 2009), and (medical) educational fields (e.g., Cheng et al., 2014). Despite the importance of debriefing to professional learning and assessment in a broad range of professions, a recent review study notes that "there are surprisingly few papers in the peer-reviewed literature to illustrate how to debrief, how to teach or learn to debrief, what methods of debriefing exists and how effective they are at achieving learning objectives and goals" (Fanning & Gaba, 2007, p. 115). Other review articles suggest that the theoretical and empirical human factors literature concerning debriefing is sparse and fragmented (e.g., Tannenbaum & Cerasoli, 2013); and the problems of debriefing tend to be attributed to implementation (e.g., Zuckerman et al., 2012) without regards to the contextual particulars.

Debriefing, which has its origin in the military, constitutes a means to learn from experience for the purpose of developing new strategies by reviewing, analyzing, and discussing pertinent (real or simulated) events (Dennehy, Sims, & Collins, 1998). Debriefing is useful especially in fields with high-stakes environments, where errors can have considerable, often deadly consequences, including combat, surgery, and aviation (e.g., Zuckerman et al., 2012). Recent reviews of the scholarly literature suggest shortcomings in the topics researched, paucity of related theory, limitations in the number of empirical studies, and problems in research design (e.g., Adler et al., 2009; Fanning & Gaba, 2007; Tennenbaum, 2013). There appears to be general agreement that debriefing is beneficial, as indicated by learning and performance improvement, which one meta-analysis suggests to be about 25% and equivalent to an average effect size of $d = 0.67$ (Tannenbaum & Cerasoli, 2013). Simultaneously, there also is agreement of considerable variation of the findings of studies included in meta-analyses (e.g., Cheng et al. 2014). Different meta-analyses consistently report that the outcomes of video-mediated debriefings are not different from debriefings without video (e.g., Cheng et al., 2014; Tannenbaum & Cerusoli, 2013). However, analyses how the video influences debriefing practices and the effects it might have on cognition could not be identified in the current endeavor.

There exists one study of debriefing in aviation (Dismukes, McDonnell, & Jobe, 2000). The study suggests that the instructor pilots with two-member crews tended to talk more (61%) than the crewmembers taken together (39%), often asking questions, with little within-crew talk. First officers (19%) and captains (20%) contributed about the same amount of talk. Considerable variance exists between airlines as to the number of words concerning crew resource management (between 19–64% of instructor pilot talk; 25–68% of crew talk) and technical topics (between 8.1–38% of instructor pilot talk; 5.6–23% of crew talk). The remaining talk was classified as mixed and non-specific. On average, 41% of the instructor pilot talk and 52% of the crew talk concerned crew management and technical performance of the crew. Finally, the debriefing lasted on average 31 minutes, with a range of an order of magnitude between the shortest and the longest meetings, which is almost twice the 17.85-min duration that a recent meta-analysis reported for all the 111 studies it included (Tannenbaum & Cerasoli, 2013).

Consistent with the assumption that language merely is a medium for externalizing thought (Beattie & Shovelton, 1999), existing studies of debriefing focus on what is said during these meetings, the number of words and the content (e.g., Dismukes, McDonnell, & Jobe, 2000). But knowing and remembering do not only exist in the form or representations but exist in implicit forms often characterized by such adjectives as *embodied, sensori-motor,* or *kinaesthetic* (e.g., Sheets-Johnstone, 2011). This has led scholars generally (e.g., Hanks, 1992) and aviation-related scholars specifically (Hutchins & Palen, 1992) to recognize communication as a distributed phenomenon covering physical space, gestures, and words. The present study therefore investigated communication more broadly, including gesture production specifically.

Pilots and flight examiners do not just communicate to externalize the contents of their minds. Instead, language and the organization of debriefing meetings are *cultural* practices. There is now a vast literature on the dependence of cognition on culture (e.g., D’Andrade, 1995; Hutchins, 1995a). In fact, sociocultural theories presuppose that every higher cognitive (psychological) function was a social relation first, and, therefore, is a cultural phenomenon (Leont’ev, 1978; Luria, 1973; Vygotsky, 1989). To properly understand cognition and cognitive development, culture needs to be taken into account (e.g., Saxe, 1991). The study of culture is the domain of anthropology. The interdisciplinary approach investigating cognition and culture simultaneously is referred to as cognitive anthropology (D’Andrade, 1995).

The review of the literature reveals the need to better understand what participants do in debriefing meetings (i.e., their cultural practices), how these are organized, how knowledge is presented, and how debriefing may lead to further knowledge (learning). This study was designed to investigate the cultural practices (patterned actions) and cognition (cultural schemas, and understandings) of debriefing in aviation. The following two research questions were investigated using both quantitative methods typical of cognition research and qualitative-descriptive methods typical of anthropology studying cultural practices:

1. How are debriefing meetings organized with respect to duration, amount of talk, amount of gesturing, and relative examiner/examinee patterns of participation?
2. How do the organization of debriefing meetings and the tools used mediate the pilots’ learning opportunities?

RESEARCH METHODS

Design

Observations, formal and informal interviews, recorded debriefing sessions, observed and recorded simulator sessions, and stimulated recall with selected flight examiners were employed in this study. The participating airlines had been invited based on their fit in a 2 (use or not of a specific pilot performance model) x 2 (use or not of a debriefing tool) factorial design.

Model of Assessment of Pilot Performance. The first factor was use or non-use of the Model of Assessment of Pilot Performance (Mavin, Roth, & Dekker, 2013). The model comprises 2 technical (aircraft flown within tolerances, aviation knowledge) and 4 non-technical performance areas (situational awareness, decision-making, management, communication) assembled in a hierarchy of enabling and essential skills (Figure 1). The airlines using this model also employ an assessment metric that allows flight examiners to evaluate pilots rating them from 1 (*unsatisfactory*) to 5 (*very good*) pilots on the 20 subcategories that the model comprises (Mavin et al., 2013). Although all flight examiners use the assessment metric for assigning grades, the use of the model or the metric as part of the debriefing sessions is diverse. In some sessions, neither tool is used; in other sessions, either the model or the metric may be used to point pilots to the categories in which they did well/poorly, and in explaining how this poor performance mitigated their performance on other human factors categories.

«««««« Insert Figure 1 about here»»»»»»

Debriefing tool. The second factor was use or non-use of the debriefing tool. The debriefing tool is an integrated system representing various aspects of the simulator session (Figure 2). It includes (a) a video of the pilots from behind (in the way the flight examiner would see them) and the view the pilots would have outside their window (alternatively, the aircraft from behind); (b) photographic representations of the flight instruments, engine instruments, flight controls, and engine controls; (c) graphical representations of altitude, speed, and vertical speed and overhead view of aircraft and programmed global navigation satellite system; and (d) a control panel for playback. During the simulator session, the flight examiner places marks that subsequently permit rapid access to selected episodes during the debriefing meeting. Depending on the way in which the debriefing session unfolds, the flight examiners select a small number from among those episodes previously marked.

«««««« Insert Figure 2 about here»»»»»»

Participating Airlines, Pilots, and Flight Examiners

Airlines. In this study a total of 29 debriefing sessions were videotaped involving five airlines operating in two countries of the southern hemisphere. All participants were selected randomly from among those available from the roster and willing to participate during the data collection periods.

Participants. The distribution of participating pilots and examiners across airlines is provided in Table 1. Characteristics for the three types of participants—pilots assessed, flight examiners in debriefing, and the subset of flight examiners in stimulated recall sessions are provided in Table 2. The six flight examiners who also participated in stimulated recall sessions came from airlines A (*n* = 2), B (*n* = 1), and D (*n* = 3).

«««««« Insert Table 1 about here»»»»»»

«««««« Insert Table 2 about here»»»»»»

Ethics protocols. The studies were approved by the university ethics committee, the companies involved, and the respective labor unions. During recruitment, all pilots were assured that their non/participation would not affect their employment and that they were free to leave the study at any time.

Contexts and Tasks

Besides the general fieldwork within the participating airlines, involving many informal exchanges with key informants such as training managers, information was collected in four task settings.

Debriefing sessions following simulator exercises. Each training/examination day started with a 1-hr brief. Pilots and examiner then entered the simulator for a 4-hr session, which was shortened by a break that tended to last between 15–20 minutes. The debriefing sessions, which were part of the participating companies' regular training and assessment regimes, concluded the day. These sessions constitute naturalistic settings and were conducted according to the company procedures. A 1-hr slot was available for each session.

Interviews associated with the debriefing sessions. All participants in the debriefing sessions were interviewed individually on three occasions: during the halfway break and at the end of the simulator exercise and following the debriefing session. During the first two interviews, pilots were asked what they remembered to have been significant and what they thought the flight examiner might bring up during debriefing. Flight examiners were asked what stood out for them and what they considered talking about during debriefing. Following debriefing, all participants were asked whether the session had unfolded as anticipated, whether there had been surprises, and how to improve on the practice of debriefing.

Flight examiner stimulated recall sessions. For these sessions, the flight examiners were shown randomly selected clips of the debriefing sessions they had conducted on that day. The clips were selected at random because their primary purpose was to make the debriefing present again and, thereby, assist the flight examiners in their recall. All examiners were asked to respond to the same set of questions. How did you prepare for the session? When did you make your decisions about the assessments for the pilots? When and how did you make the decisions about the particular details to be discussed during the debriefing meeting?

Data Collection

General observations. General observations were recorded in field notebooks and by means of digital cameras, which also were used in place of photocopiers for recording the nature and contents of artifacts (e.g., white board contents, artifacts used in sessions, or physical contexts). Handwritten field notes often were elaborated in electronically kept notes at the end of a day in the field. The database includes the manufacturers' systems manual, the manufacturers' general and company-specific standard operating procedures, copies of the quick reference handbooks, landing plates for the airports involved, charts, speed books, and other materials useful or required for analyzing the talk about specific simulator events.

Debriefing sessions. The debriefing sessions were videotaped generally using three cameras. Two digital cameras were positioned in opposite corners of the debriefing room so that all aspects of the sessions were captured, including use of artifacts placed on one of the three walls of the rooms. The built-in camera of a laptop computer was used as a backup. In addition, all sessions were audiotaped with a digital recorder for high quality sound and rough transcription purposes. A total of 17:44 hours of debriefing were recorded.

Interviews associated with debriefing sessions. The three interviews per pilot and flight examiner surrounding the debriefing session were audiotaped. A total of 16:27 hours of interviews were recorded, with a mean of 4.04 minutes per interview.

Flight examiner stimulated recall sessions. The stimulated recall sessions were videotaped using one camera seated behind the flight examiners. The frame included the video shown to stimulate recall and any artifacts used, such as the flight examiners' personal notebooks, which they had been asked to bring to the sessions. A total of 5:05 hours were recorded, yielding a mean of 50.8 minutes per session.

Transcription. All recordings were transcribed verbatim in their entirety by a professional service employing a transcriber who has piloting experience. To ascertain accuracy, the transcriptions were reviewed and compared to the recordings in their entirety .

Data Analyses

Videotapes were analyzed in sessions with colleagues based on the precepts of *interaction analysis*, a method privileged by those interested in the study of cognition at work (Jordan & Henderson, 1995). It involves viewing videotapes to identify invariants and structures of cultural behavior—e.g., the role of semantic and episodic memory and the structural organization of debriefing meetings—through the interactions among colleagues. Any coding scheme that evolved was discussed and tested jointly to ascertain consistency. For the quantitative modeling, standard, frequency-based statistical procedures are used (e.g., *t*- and *F*-tests). Potential limitations of this study arise from the small number of cases for cells in the *F*-tests, which limit statistical power and the possibility to detect true differences. Where appropriate and doable with available software, Bayesian tests were used because these afford evaluating the relative probabilities of null and alternative hypothesis given the data (Rouder, Speckman, Sun, & Morey, 2009).

FINDINGS

In the following two sections, answers are provided to the research questions by means of assertions that are subsequently substantiated with evidence from the database.

Broad Descriptors of the Cultural Practices of Debriefing

Research Question 1. How are debriefing meetings organized with respect to duration, amount of talk, and relative examiner/examinee patterns of participation?

Assertion. (a) Debriefing meetings differ in terms of the duration of sessions across airlines; (b) debriefing sessions differ in terms of amount of flight examiner talk but not in pilot talk; (c) flight examiners talk significantly more than pilots; and (d) companies differ with regard to the timing of assessment results.

Duration. Previous research suggests that the line-oriented flight training (LOFT) debriefing sessions tend to last about 30 minutes (Dismukes et al., 2000). There was a suggestion that longer debriefing sessions were desirable. In the present study, there was a 1-hr slot available for the debriefing session in all airlines, including the completion of the paperwork. The debriefing meetings lasted 36.4 minutes (SD = 14.3) on average. However, there was a statistical significant effect ($F(3,23) = 17.24$, $p < .0001$) in the duration of the sessions by company. A Tukey HSD test reveals the statistical significance of five of these differences (Table 3). In airline D, there was one session that was less than half as long (23.2 minutes) than the mean of the 10 remaining sessions (48.3 minutes). This session focused on assessment only. The two pilots were very experienced. Even though it ended with the pilot flying in the captain's seat received a 2 (out of 5), *minimum standards*, on situation awareness because of two incorrect turns, there appeared to be agreement that everyone had known, and learned from, what had happened. The shorter than normal session in airline C—30.5 minutes compared to a mean of 45.6 minutes for the remaining sessions—occurred in the context of the non-functioning of the debriefing tool so that none of the marked episodes could be replayed. The sessions of the two companies with a debriefing tool lasted longer, as expected, due to the (a) additional time spent on actually viewing video recordings and (b) additional opportunities that arose for analyzing the behavior of the pilots seen on video. The amount of time spent may be affected also by the timing of the sessions.

Thus, when examinations are taking place in the middle of the night with a debrief set for 2am in the morning, flight examiners tend to be conscious of not making the session last too long.

«««««« Insert Table 3 about here »»»»»»

Amount of talk. Time alone is not inherently the best quality indicator of debriefing sessions. Instead, how much was said to account for and assess performance may constitute a better overall measure. If the number of words per session are used as dependent variable, there also is a statistically significant effect ($F(3,23) = 11.61$, $p < .0001$). However, the HSD test shows only three of the differences to be significant: A vs. B, B vs. C, and B vs. D (Table 4). Surprisingly, a comparison of the number of words contributed by the pilots did not differ across the companies ($F(3,24) = 1.36$, $p > .05$). That is, across the different airlines, the pilots contributed about the same number of words and, therefore, to the making present of the preceding simulator experience. The effect of different number of words, therefore, mainly arose from the differential amount of talk on the part of the flight examiners.

«««««« Insert Table 4 about here »»»»»»

Participation. In educational research, lecturing and other instructor-centered teaching strategies not only are experienced as boring but also range among passive learning methods that tend to be inconsistent with the most recent learning theories; active learning methods that encourage active learner participation in dialogue tend to be more efficacious especially in the training of practitioners (e.g., Rogal & Snider, 2008). Previous work in aviation emphasizes the desirability of more pilot participation and a move toward facilitated debriefing (Dismukes et al., 2000; Dismukes & Smith, 2000). Regulators—among others in the certification of flight examiners (e.g., CAA-NZ, 2013)—also emphasize crew participation. In this study, however, the flight examiners talked significantly more ($M_{\text{FE}} = 4{,}163$ words, $SD_{\text{FE}} = 1{,}876$) than pilot pairs (M_{P} 1,223 = words, $SD_{\text{P}} = 587$) ($t(27) = 8.74$, $p < .0001$). To control for the different session durations, a one-sample *t-test* was conducted to test whether the ratio of the number of words flight examiners talked to number of words of both pilots significantly differed from $WR = 1$ (equal amount of talk of flight examiners and pilots). Flight examiners produced 4.3 times more words than the two pilots facing them taken together ($t(27) = 6.57$, $p < .0001$). A Bayesian test indicates (JZS Bayes Factor = $2.21*10^{-5}$) this to be *decisive* evidence in favor of the alternative hypothesis, which is over 45,000 times more likely than the null hypothesis of equal amount of talk. Although the omnibus *F*-test reveals a statistically significant effect between the four airlines ($F(3,24) = 3.48$, $p < .05$), the differences fail to reach statistical significance when the HSD test is used. If the effect is real, then the low statistical power due to the small number of sessions per cell may be at the origin of this phenomenon. Thus, there is considerable variation even within the airlines of the ratios of flight examiner to pilot words (Figure 3), which ranges in flight examiner to pilot word ratio $WR < 1$ to $WR > 10$, over half falling in the range $1 < WR \leq 4$. The sessions in airlines B and C fall to the left of most sessions in airline D; in both of these airlines, there had been changes recommended to the practice according to which pilots were to contribute more to the debriefing discussion that had been done before. A more detailed analysis reveals that there is considerable within-flight-examiner consistency. For those examiners observed repeatedly, one flight examiner had five sessions with word ratios $6.7 < WR < 8.99$, whereas another was observed on four occasions with ratios $2.09 \leq WR \leq 3.42$; and two others were observed in two sessions with $1.28 \leq WR \leq 1.41$ and $4.07 \leq WR <4.47$. The stimulated recall sessions, where flight examiners are shown fragments of their sessions, and interviews following initial analysis show that flight examiners who speak a lot tend not to be aware of this. For example, one flight examiner noted that the pilots "were interactive" and that he was more interactive than he had been in previous years. Yet the analysis revealed $WR = 7.36$, that is, the flight examiner had used 7.36 as many words as the two pilots combined.

«««««« Insert Figure 3 about here »»»»»»

Production of flight-related hand/arm and body movements. Important aspects of cognition, especially in technology-rich environments, are articulated by means of hand/arm and body movement (Heath & Luff, 2000; McNeill, 2000; Zemel, Koschmann, LeBaron, & Feltovich, 2008). An existing study in aviation reported a mean of $X = 82.3$ ($SD = 42.5$) flight-related hand/arm or body movements when narratives of the event were produced but only $X = 22.7$ ($SD = 17.5$) gestures when the talk was driven by an assessment metric assessment (Roth & Mavin, 2014); the sessions had lasted a mean of 65.6 minutes. Because gestures are produced in parallel with associated words to which they correspond (e.g., Hadar & Butterworth, 1997; McNeill, 1992), a linear relation between number of gestures and words should be anticipated. In the present study, too, there was a comparable number of flight-related (iconic) gestures produced per session ($X = 34.2$, $SD = 25.5$). The number of gestures produced is correlated with the number of words ($r = .667$, $p < .0001$). The addition of a quadratic contribution to the linear model increases the explained variance by a non-significant mount from $R^2 = .445$ to $R^2 = .448$ ($sr = .0055$, $p > .05$). As the plot of the data shows (Figure 4), however, in the linear approach there are three possible outliers in sessions with very low gesture frequencies. For all three sessions, data are available on the flight examiners involved. In each case, the low frequency of gestures is the exception as they normally produced 2.5 to 4 times as many gestures. If the three marked data points are indeed outliers, disregarding them would change the correlation significantly ($r = 0.834$, $p < .0001$). In both cases, Bayesian analysis shows that there is decisive evidence against the null hypothesis (alternative hypothesis is 2,900 and 9,900 more likely, respectively).[1]

«««««« Insert Figure 4 about here»»»»»»

As may be expected, there were few gestures in the sessions of company B, where the sessions had the shortest duration, and which focused mostly on evaluation with little elaboration of the experienced events and how to improve upon performance (Figure 4). The three possible outliers to the general trend are sessions where a debriefing tool was used (Figure 4). The two sessions from airline D involved the same flight examiner and crew. All three were very experienced pilots, the person in the right-hand seat a senior flight examiner himself, and the captain for 5 years in that rank.

There are no statistically detectable differences ($t(27) = 1.66$, $p > .05$) in the number of flight-related gestures per 100 words employed between flight examiners ($X_{\text{FE}} = 0.78$, $SD = 0.45$) and pilots ($X_{\text{P}} = 1.13$, $SD = 1.05$). There is, however, a statistically significant effect of the number of

[1] Although there is no theoretical reason for expecting anything other than a linear relationship between the number of gestures (g) and the number of words (w), other models were tested yielding slightly higher correlations than the uncorrected linear model: $g = 3.95 \cdot 10^{-5} \cdot w^{1.6}$: $r = .735$ ($p < .0001$).

gestures per 100 words across the four airlines (*F*(3,24) = 3.63, *p* < .05). Based on Tukey's HSD test, there is a significant difference at the α = .05 level between airline A and airline B but no differences with or between airlines C and D (Table 5). In the debriefing sessions of airline B, where little elaboration of actual flight situations was observed, the mean number of gestures per 100 words is lowest; in airline B, which had neither MAPP nor the debriefing tool, the number of gestures per 100 words was highest. The ratios were nearly identical for the two airlines with the debriefing tool.

«««««« Insert Table 5 about here »»»»»»

Timing of assessment announcement. The timing of the announcement of the overall assessment differed according to company. In airlines A, B, and E, the assessment was explicitly done at the beginning, often with a "congratulations," especially when a pilot assessed was more inexperienced or "has had some issues." The degree or intensity of mitigating aspects immediately followed. In airlines C and D, on the other hand, the result of the assessment was provided at the end—though in many instances, such as when the flight examiners had been rating each exercise discussed, the overall outcome (pass) was implicit. In two of the 29 simulator sessions, the captains in question had begun making wrong 90° turns (one captain twice turned right where the turns should have been left; one captain turned onto a 15- rather than 10-mile arc). Everyone involved knew these had been serious mistakes. During the think-aloud protocols where pairs of flight examiners evaluated the performance of a similar wrong turn, every pair had failed the captain. The errors had been serious and the pilots thought right to the end of the debriefing meeting that they had failed the examination, which would have led to being taken off-line and to having to enter a retraining schedule. It was only when the flight examiner announced that the captains would receive "their stickers" (i.e., the 12-month renewal of the pilot license) that the overall passing grade was revealed.

Organization- and Tool-Mediated Nature of Pilots' Learning Opportunities

Research Question 2. How do the organization of debriefing meetings and the tools used mediate the pilots' learning opportunities?

Assertion. Pilots, having experienced the simulator sessions as events so packed that they were exhausted, had forgotten much of the details of what they had done. The tools (model of assessment, debriefing tool) afford different conceptual organization of the meetings that differentially assist pilots in recalling events. These two aspects mediate what pilots can reflect upon and learn from. The available (cultural) tools—the human factors model of pilot performance and the debriefing tool—provide affordances for different conceptual organizations of the debriefing meetings.

Background. Past research on the practices of debriefing noted that simulator sessions constituted "busy, intense experience" (Dismukes et al., 2000, p. 35). But this research did not investigate how this characteristic may affect the structure and content of the debriefing sessions. What and how much will participants in busy and intense experience remember? In the cognitive sciences, a distinction is made between episodic and semantic memory (Tulving, 1984). Episodic memory has as information source sensations, is organized in units of events and episode, has a temporal organization, is registered experientially, is temporally coded directly, provides limited inferential capability, and is more context dependent. Semantic memory has comprehension as its source, is organized into units of facts, ideas, and concepts, is atemporal, and has rich inferential capability.

Duration and intensity of the simulator sessions mediates cognitive events in the debriefing sessions. In this database, there are many instances where pilots talk about having forgotten specific aspects of their flight (exercises) or where the debriefing talk shows that they were not aware of their actions, instrument readings, or control settings. Even when they were provided with opportunities to talk about events in the debriefing, pilots frequently did not remember critical aspects of the flight, although they might have done so when the flight examiner addressed the issue or when they saw themselves on the video included in the debriefing tool. Crewmembers more easily remembered events that required the repetition of a task, which always followed a performance rated as unsatisfactory. That is, because the exercises had to be flown again, which highlighted the unsatisfactory nature of the first attempt, associated events clearly stood out for the pilots. They had made some error that was significant enough to warrant the repeat exercise. It was the stopping of the simulator session, caused by something that has happened, followed by the repeat of the exercise, that made the event stand out from the inchoate stream of experience, in part because of the affective qualities associated with failure.

In addition to the recall problem under normal condition, simulator sessions also take place late at night and in the early morning hours. Debriefing following such sessions are even more difficult and exhausted pilots tend to have greater difficulties remembering what actually had happened. Flight examiners know the effect of the simulator sessions on the pilots and adapt how they conduct the debriefing and its duration. Especially when they had assessed a particular performance as requiring a repeat of the exercise, flight examiners debriefed what had happened in the simulator and then conducted the repeat of the exercise. In fact, although company policy may suggest the repeat to be conducted at the end of the simulator session, some flight examiners conducted the debriefing immediately. Pilots tend to find it helpful when examiners review problems immediately, as the learning opportunities are made available right then and there when the situation really is present to them in vivid detail. The practice therefore supports cognition and learning in conditions characterized by considerable intensity and tremendous fatigue. Less experienced first officers and captains are more prone to forgetting than and more experienced ones, especially those who serve as training captains and flight examiners—often associated with lower situation awareness during the exercises. The most experienced pilots, often flight examiners themselves (n = 6 in 10 sessions), tended to be keenly aware of all aspects of the flight and aware of smaller details, and could articulate reasons why they had acted in the way they did rather than implementing other possible ways.

The tools mediate the cognitive organization of debriefing sessions: episodic and chronological order. Although the related research suggests that episodic memory is more vulnerable than semantic memory, in four companies (A, C, D, E) the debriefings were in the order of the tasks or in the chronological order. Using their notes, moving from the beginning to the end, flight examiners addressed any issues that they had identified and, following the simulator session and their constitution of the overall assessment, had marked as needing to be addressed. The notes served as external memory devices with sequential access, functioning much like early computers with tape as storage device. Even though pilots were sometimes invited to reflect on their performance (especially in airline A), which could involve any part of the preceding exercise, the chronological order of their occurrence and the linear order of the written notes then drove the sessions. The flight examiners described relevant events in considerable detail and highlighted, as pertinent, positive or negative performance aspects. Flight examiners moved in this way from event to event until the entire 4-hour simulator sessions had been covered. At that point, the final assessment of an often-implicit passing grade was announced or reiterated thereby ending the debriefing; this was followed, when relevant, by the signing of documents. In airlines C and D, replaying such events using the debriefing tool provided pilots with opportunities to see what they were doing or for all parties to verify a verbal description, especially when examiners and pilots provided different versions.

In airline B, the sessions were organized according the conceptual model (MAPP), which, in laminated form, was placed on the desk between pilots and flight examiners. In this airline, the training managers had implemented the model of pilot performance, associated with the ways in which flight examiners were asked to reorganize their debriefing sessions according to the main categories of the model. This was to replace the chronological order that had worked for them because, as one of them said, “that’s how I play [the session] back in my mind, start to finish.” Rather then covering the preceding simulator sessions in their chronological order, the flight examiners tended to (a) point to performance categories—usually “enabling skills” before “essential skills” (Figure 1)—and (b) then talk about a number of events where strengths and weaknesses of a particular skill were apparent. Riffling through their notes, where they had marked their observations using the first letters of the performance categories from the model, the flight examiners in this airline then referred to other events consistent with the assessment or a lapse that they had observed. Here, the notes served as external memory devices with random access typical of computers with hard drives. In this manner, the flight examiner covered the other performance categories in the lower part of the performance model and how these affected those categories in the upper part. Once the three categories were covered, the debriefing was completed.

In airline D, although it used the same conceptual performance model as airline B, the debriefing meetings were organized according to the chronological order of events and the linear order of the notes. However, following each event, examiners tended to make their assessments in terms of the conceptual performance model. In some instances, flight examiners not only assessed the different components but also had a laminated version of the assessment metric. For example, in one instance the flight examiner pointed to the knowledge category in the assessment metric oriented so that the pilots could see it right side up, and read back to them the "word pictures" that go with assessment scores 5 and 4. He then identified the pilots as having performed at a level corresponding with the word picture under score 4. Each event was assessed in terms of the assessment metric that the company uses, thereby providing the pilots with an understanding of how well they had performed and, depending on the circumstances, why a lower performance in one area (e.g., communication) improved or decreased performance in another area.

DISCUSSION

This study was designed to investigate the cultures and cognition of debriefing in the aviation industry. A consistent shortcoming of all studies and meta-analyses reviewed lies in the fact that outcome measures do not provide information on the debriefing as unfolding process, always adjusting itself to the contingencies of the nature of the simulator event, which never is the same even when the overall exercise is the same for all crews involved. Even when different pairs of flight examiners discuss and evaluate the same event, there are considerable differences in what they pick out to be salient, which technical or non-technical factor is at issue, and which facts are pertinent to a crew's success or failure (e.g., Roth et al., 2014a).

A previous study on debriefing in aviation investigated external factors (Dismukes et al., 2000). The present study, focusing on the internal dynamics of debriefing sessions, shows that there are cognitive consequences of cultural aspects such as the duration of debriefing meetings, the amount of talk and who talks, and the debriefing culture. Thus, the duration of debriefing and number of words are proxy measures for the number and detail of significant events discussed. Because of the correlation with the number of flight related gestures, the duration also affects the amount of the original events that can be made present again and, therefore, how much of past flight-related cognition is present for reflection. The significant asymmetry between flight examiner and pilot talk changes the relationship between what is verbally described and visually enacted for the pilots versus what and how much of their *embodied* knowing they themselves make present through enacted gesture sequences. The power differential between flight examiners and pilots also has consequences for the assessment of the degree to which the pilots' cognitive performances constitute the norm versus constituting deviations from the norm. The power differential mitigates any more-symmetrical approach to debriefing, such as the *facilitated debrief,* where the roles taken are more symmetrical (Dismukes et al., 2000).

Initial ethnographic observations suggested that the human factors based assessment model (MAPP, Figure 1) might lead to a reorganization of the assessment exercise. This study does not provide evidence for a single main effect, as there were different patterns observed in the airlines using this tool. In airline B, the debriefing meetings were organized—in order and content—according to the structure of the tool. In airlines D and E, on the other hand, an episodic (chronological) organization was observed. The debriefing tool (Figure 2) supported pilots in remembering what had happened. Although the debriefing tool provides random access to episodes from the simulator session, the debriefing sessions in airline D where both tools are used still had an overall episodic organization.

Debriefing sessions often are structured episodically. The temporal organization of debriefing is easily supported by the linearity of the notes taken during the session and is associated with great apparent detail that become available as soon as the flight examiners begin to narrate and enact some event—a degree of detail much greater than any of the notes they have made. In those sessions organized semantically based on major conceptual performance categories, the relevant details that a sound analysis of the events became available to the participants with episodic accounting of what has happened. The conceptualization of the differences, however, focused on the two forms as propositional memory, whereas this study shows that the movement and event sequences are not propositional but associated with bodily knowing. The movements observed are not symbolic *re*presentations but are the same movements that get the work done and orient in the work environment generally and in the aircraft cockpit specifically (Roth et al., 2014b).

The predominant order of topics was a function of the episodic order of the events, and the linear organization of the flight examiners' notes. Thus, the practice drives the order of representation and, therefore, the associated cognitive efforts. The order of observations in the representational tool drives the flight examiners' cognitive organization of what has happened and how it is to be evaluated. The order also contributes to how they organize debriefing practice. There is no inherent necessity to discuss the simulator sessions in the chronological fashion in which they had unfolded. However, the notes written during the sessions are organized in a linear fashion, and flight examiners tended to work through events and associated notes from the beginning to the end. The debriefing tool, even though it affords accessing the marked events in a random-access fashion, was used in a sequential way. In fact, the one time a pilot did request such (random) access, the flight examiner, suggesting that this would take time, continued following his predetermined agenda.

When flight examiners used a semantic conceptual ordering, the practice was associated with the production of less-detailed accounts of what actually happened and more concerned with the articulation of the conceptual categories. However, the less-experienced pilots and pilots in training felt that they might have gotten more out of the sessions if these had covered the events in greater detail. The previous study in the field (Dismukes et al., 2000) looked at full flight exercises only, but debriefing sessions may be affected by factors such as precisely what the purpose of the session is and what the contexts are. Thus, in a "spot check," there are many brief exercises, which put much higher demands on pilots' memory and, therefore, on opportunities for learning.

One cannot talk about the debriefing sessions without articulating the particulars of the contexts within which they take place. Simulator training and assessment are much longer than the line flights that these regional pilots conduct and, because the former are packed either with exercises or high workload events, they are physically and affectively draining the participants. This mitigates what and how much pilots remember to have happened. Existing research places primacy on crew-driven debriefing sessions (e.g., Dismukes & Smith, 2000). However, if the crew has problems remembering some or all of the relevant detail, then a crew-driven approach alone cannot be the answer. A significant amount of the debriefing sessions needs to be made to make the past experience present again with sufficient detail so that debriefing it will be associated with learning. It is the debriefing process generally and the flight examiner specifically that play an important role in producing the content for the reflective process. The cultural practices reproduced the flight examiners' hold over facts and knowledge, thereby establishing what there was to be learned and how. This influenced (a) the cognition shared by participants present in the debriefing room and (b) what pilots could "take-away," that is, what would shape their future cognition.

In this study, there was considerable variation about the relative amount of talk that the flight examiners and crewmembers contributed to the debriefing meetings. Past research suggested the desirability of increase in the amount of crewmember talk (Dismukes et al., 2000). The present study shows that pilots tend to be aware that they forget much of the detail. This influences what they want from the debriefing session and affects their form of participation. Thus, especially less experienced pilots found it more useful to have the flight examiner narrate back to them what they had done and tell them why it had been wrong and how to improve upon it. In such situations, pilots find it less useful to take a more proactive role and debrief the simulator session themselves. This study also shows that the awareness of having forgotten detail contributes to the relevant authority over just what the factual performance had been, and, therefore, how it was to be evaluated.

Debriefing can only be useful if the pilots actually remember what they have done, recognize its variance from best practice, and can actively work to change it. Because much of the cognition in an aircraft cockpit is situated and distributed (e.g., Henriqsen et al., 2011; Hutchins, 1995), the presence of past experience likely is affected by the extent to which pilots and flight examiners can and do act out what has happened. In the present study, there was a linear correlation between the amount of talk and the number of flight-related gestures and body movements. These flight-related hand, arm, and body movements serve a symbolic function; yet their accuracy in terms of locating the instrument or actuator relative to the body-centered system is amazingly high. Thus, an analyst familiar with the aircraft can easily identify what someone is talking about even when the soundtrack is turned off because the movements are not merely generic but reflect the physical layout of the cockpit. There are characteristic hand movements and word exchange sequences that make recognizable overall flows or bodily *kinetic melodies* (Roth et al., 2014b). These flow sequences are not encoded in terms of their parts but unfold on their own once triggered.

The results of this study provide evidence for the interaction of cognition and culture. Thus, for example, the cultural practice of debriefing depends, in its structure, on the cognitive organization of the flight examiner notes and episodic memory; but the particular order of the notes and episodic memory are determined by the cultural practices on which examinations are built. On the part of the pilots, what they remember is a prerequisite for their subsequent learning (Ausubel, 1968), but the cultural practices of the examinations mitigate what and the extent to which pilots can remember and subsequently analyze during the meetings. Consistent with these interactions of cognition and culture, this study draws on cognitive anthropology as method. The benefits are apparent when the present study is compared to the results of a previous one (Dismukes et al., 2000), which was oblivious to important cultural factors that mediate the particular outcomes of empirical (statistical) approaches.

IMPLICATIONS

This study already has led to changes in three of the participating airlines, where the second author has led workshops for flight examiners. Together with the training managers of the airlines, a five-phase approach to debriefing resulted from the empirical study. In phase 1, the flight examiner invites the pilots to review the plan that was established prior to the simulator session. In phase 2, the flight examiner invites the pilots to outline the strengths and weaknesses of their simulator performance. In phase 3, the pilots are encouraged to relive specific events by talking them through to the fullest extent possible, including the use of the video that recorded the entire session. In phase 4, the pilots analyze what went right and what went wrong, an endeavor in which they are assisted by the flight examiner. As a result of phase 4, pilots may return to phase 3 for further narrative articulation of their experience. Finally, in phase 5, the pilots and flight examiner review the simulator session and articulate the “take-home” points. At the time of this writing, a new study is in the planning stage designed to investigate the impact of this new debriefing practice on pilot cognition and culture of debriefing.

Early data of a study in airline D currently in progress show that there has been a decrease in the use of the debriefing tool. Simultaneously, the flight examiners suggest that pilots with problematic performances—i.e., who score low on the company’s assessment metric based on the human factors model (Figure 1)—also have difficulties remembering past events. The use of the debriefing tool might assist such pilots in reflecting on their experience. A new study is currently designed to investigate the relationship between performance level, problems of recall, and the mediating roles of debriefing procedures and the debriefing tool.

Flight examiners who use the debriefing tool know that it assists pilots in making present again what had happened. But it is not only the video that can produce this effect. Pilots tend to have similar experiences when the debriefing allows them to articulate *themselves* in detail what they have experienced. Acting out and talking through flow sequences makes present again, for everyone to see, what the pilots have done or what they should have done. In contrast, when the flight examiner simply describes events verbally, pilots may not know what he is talking about, leading to a “sort of foggy look in their eyes.” One line of future research might investigate how particular representational tools support pilots in making previous experience present again, via re-enactment of embodied knowing and event sequences, by means of presentations (e.g., maps, instruments, or on-board manuals), or facilitated by representations (e.g., video, graphs of speed, flight level).

ACKNOWLEDGMENTS

This study was supported by various grants from the Griffith University and Industry Collaborative Scheme. I thank the airlines, pilots, and flight examiners for their participation. My special thanks go to Tim Mavin for his contributions to the recruitment of the airlines and data collection; he provided comments on the first draft and a description of how the recommendations were implemented in further training.

REFERENCES

Adler, A.B., Bliese, P.D., McGurk, D., Hoge, C.W., & Castro, C.A. (2009). Battlemind debriefing and battlemind training as early interventions with soldiers returning from Iraq: Randomization by platoon. *Journal of Consulting and Clinical Psychology, 77,* 928–940.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart, & Winston.

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica, 123*, 1–30.

Cheng, A., Eppich, W., Grant, V., Sherbino, V., Zendejas, B., & Cook, D.A. (2014). Debriefing for technology-enhanced simulation: A systematic review and meta-analysis. *Medical Education, 48,* 657–666.

Civil Aviation Authority of New Zealand (CAA-NZ). (2013, February). Flight test standards guide: Airline flight examiner rating. Accessed August 20, 2013 at http://www.caa.govt.nz/pilots/Instructors/FTSG_Airline_Flt_Examiner.pdf

D'Andrade, R. (1995). *The development of cognitive anthropology*. Cambridge, UK: Cambridge University Press.

Dennehy, R.F., Sims, R.R., & Collins, H.E. (2009). Debriefing experiential learning exercises: A theoretical and practical guide for success. *Journal of Management Education, 22,* 9–25.

Dismukes, R.K., McDonnell, L.K., & Jobe, K.K. (2000). Facilitating LOFT debriefings: Instructor techniques and crew participation. *International Journal of Aviation Psychology, 10,* 35–57.

Dismukes, R.K., & Smith, G.M. (Eds.). (2000). *Facilitation and debriefing in aviation training and operations*. Aldershot, UK: Ashgate.

Fanning, R.M., & Gaba, D.M. (2007). The role of debriefing in simulation-based learning. *Simulation in Healthcare, 2,* 115–125.

Hadar, U., & Butterworth, B. (1997). Iconic gestures, imagery, and word retrieval in speech. *Semiotica, 115,* 147–172.

Hanks, W.F. (1992). The indexical ground of deictic reference. In A. Duranti & C. Goodwin (Eds.), *Rethinking context: Language as an interactive phenomenon* (pp. 43–76). Cambridge: Cambridge University Press.

Heath, C., & Luff, P. (2000). *Technology in action*. Cambridge, UK: Cambridge University Press.

Henriqson, E., van Winsen, R., Saurin, T. A., & Dekker, S.W.A. (2011). How a cockpit calculates its speeds and why errors while doing this are so hard to detect. *Cognition, Technology, and Work, 13*, 217–231.

Hutchins, E. (1995a). *Cognition in the wild*. Cambridge, MA: MIT Press.

Hutchins, E. (1995b). How a cockpit remembers its speeds. *Cognitive Science, 19*, 265–288.

Hutchins, E., & Palen, L. (1997). Constructing meaning from space, gesture and speech. In L.B. Resnick, R. Saljö, C. Pontecorvo, & B. Burge (Eds.), *Discourse, tools, and reasoning: Essays on situated cognition* (pp. 23–40). Berlin, Germany: Springer.

Leont'ev, A. N. (1978). *Activity, consciousness and personality*. Englewood Cliffs, NJ: Prentice Hall.

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*, 53–62.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: University of Chicago Press.

McNeill, D. (2000). *Language and gesture*. Cambridge, UK: Cambridge University Press.

Morrison, J.E., & Meliza, L.L. (1999). *Foundations of the after action review process* (Special Report 42). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Rogal, S. M. M., & Snider, P. D. (2008). Rethinking the lecture: The application of problem based learning methods to atypical contexts. *Nurse Education in Practice, 8*, 213–219.

Roth, W.-M., & Mavin, T.J. (2014). Peer assessment of aviation performance: Inconsistent for good reasons. *Cognitive Science*. DOI: 10.1111/cogs.12152

Roth, W.-M., Mavin, T.J., & Munro, I. (2014a). Good reasons for high variance (low interrater reliability) in performance assessment: A case study from aviation. *International Journal of Industrial Ergonomics , 44*, 685–696. DOI: 10.1016/j.ergon.2014.07.004

Roth, W.-M., Mavin, T.J., & Munro, I. (2014b). How a cockpit forgets speeds (and speed-related events): toward a kinetic description of joint cognitive systems. *Cognition, Technology and Work*. DOI: 10.1007/s10111-014-0292-0

Rouder, J.N., Speckman, P.K., Sun, D., & Morrey, R.D. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

Saxe, G. B. (1991). *Culture and cognitive development: Studies in mathematical understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sheets-Johnstone, M. (2011). *The primacy of movement* (2nd ed.). Amsterdam, The Netherlands: John Benjamins.

Tannenbaum, S.I., & Cerasoli, C.P. (2013). Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors, 55,* 231–245.

Tulving, E. (1984). Précis of *Elements of episodic memory*. *Behavioral and Brain Sciences, 7,* 223–268.

Vygotsky, L. S. (1989). Concrete human psychology. *Soviet Psychology, 27*(2), 53–77.

Zemel, A., Koschmann, T., LeBaron, C., & Feltovich, P. (2008). “What are we missing?” Usability’s indexical ground. *Computer Supported Cooperative Work, 17*, 63–85.

Zuckerman, S.K., France, D.J., Green, C., Leming-Lee, S., Anders, S., & Mocco, J. (2012). Surgical debriefing: A reliable roadmap to completing the patient safety cycle. *Neurosurgical Focus, 33*(5): E5, 1–8. DOI: 10.3171/2012.8.FOCUS12248

Table 1. Design of this study

		Debriefing Tool	
		No	Yes
MAPP	No	**Airline A** (turboprop 1) (*n* = 5 sessions: 10 pilots, 4 flight examiners)	**Airline C** (turboprop 2) (*n* = 6 sessions: 6 pilots, 3 flight examiners)
	Yes	**Airline B** (turboprop 1) (*n* = 6 sessions: 10 pilots, 3 flight examiners) **Airline E** (jet) (*n* = 1 session: 2 pilots; 1 flight examiner)	**Airline** D (turboprop 2) (*n* = 11 sessions: 10 pilots, 4 flight examiners)

Table 2. Characteristics of the three types of participants

N	Age	Years Piloting	Flying Hours	Examining
(Gender)	(SD)	mean	Mean	Years
		(SD)	(SD)	(SD)
		((Range))	((Range))	((Range))
Pilots				
38	37.2	12.7	5,710	
(3 f, 35 m)	(8.3)	(8.1)	(3,431)	
		((4–34))	((1,200–16,000))	
Flight examiners (debriefing)				
15	48.8	27.1	13,210	9.8
(15 m)	(8.8)	(9.3)	(3,969)	(7.5)
		((14–45))	((7,400–22,000))	((0.8–23))
Flight examiners (stimulated recall)				
6	44.7	24.3	11,900	7.2
(6 m)	(6.8)	(5.9)	(3,590)	(6.1)
		((19–34))	((7,400–17,000))	((0.8–18))

Table 3. Differences of mean duration of debriefing in four airlines and statistical significance based on the Tukey HSD test

	B	C	D(MAPP/DT) *M* = 46.1, *SD* = 9.2
A (nMAPP/nDT) *M* = 31.3, SD = 13.8	-14.8*	14.3*	14.8*
B (MAPP/nDT) *M* = 16.5, *SD* = 4.0		29.1**	29.6**
C (nMAPP/DT) *M* = 45.6, *SD* = 7.3			

* *p* < .05, ** *p* < .01

Table 4. Differences of mean number of words in four airlines and statistical significance based on the Tukey HSD test

	B	C	D(MAPP/DT) *M* = 6424, *SD* = 1,211
A (nMAPP/nDT) *M* = 5652, SD = 2,371	-3,016**		
B (MAPP/nDT) *M* = 2,636, *SD* = 687		3,607**	-3,787**
C (nMAPP/DT) *M* = 6,244, *SD* = 1,211			

** *p* < .01

Table 5. Differences of mean gestures per 100 words in four airlines and statistical significance based on the Tukey HSD test

	B	C	D(MAPP/DT) *M* = 0.60, *SD* = 0.41
A (nMAPP/nDT) *M* = 0.93, SD = 0.20	-0.63*		
B (MAPP/nDT) *M* = 0.30, *SD* = 0.19			
C (nMAPP/DT) *M* = 0.61, *SD* = 0.28			

* *p* < .05

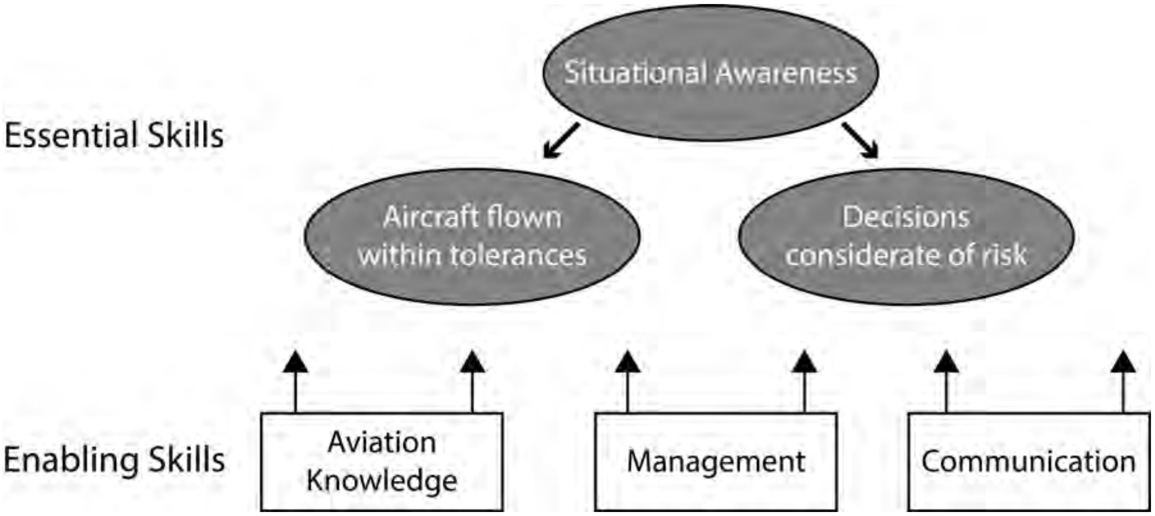

Figure 1. Two companies used the *Model of Assessment of Pilot Performance* as part of their professional development and assessment practices.

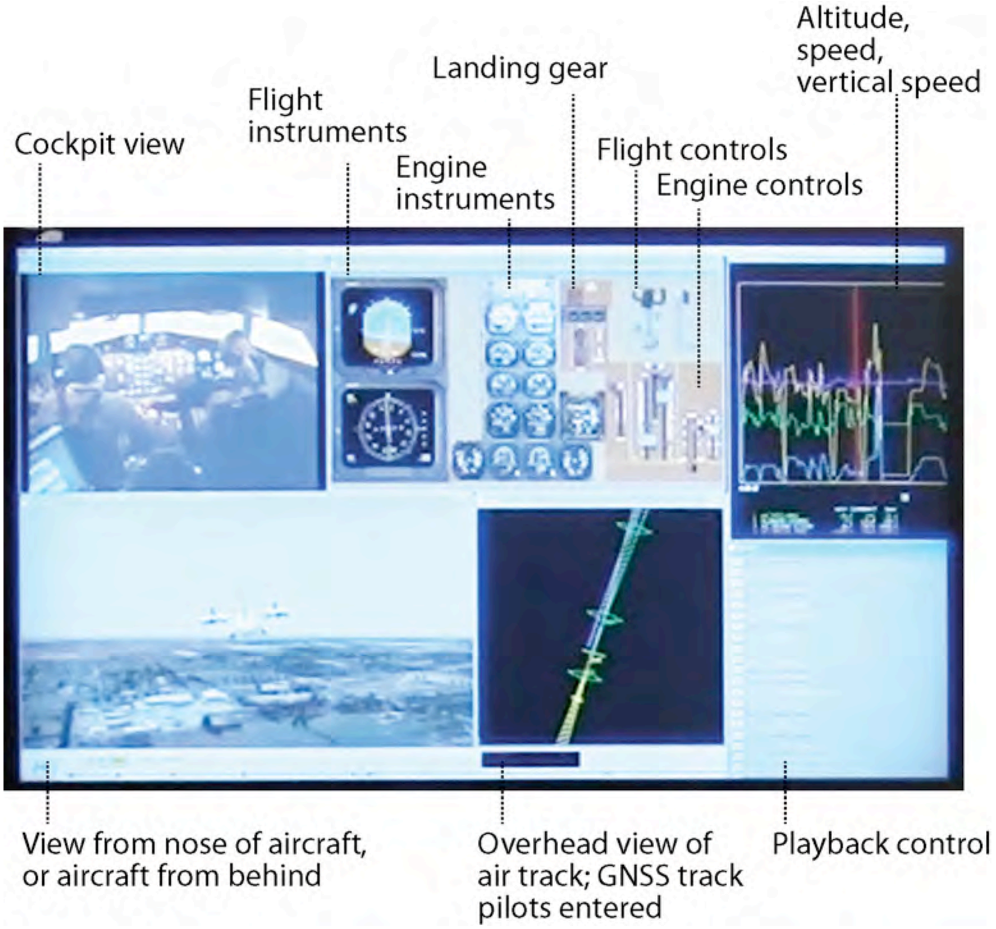
Cockpit view
Flight instruments
Landing gear
Engine instruments
Flight controls
Engine controls
Altitude, speed, vertical speed
View from nose of aircraft, or aircraft from behind
Overhead view of air track; GNSS track pilots entered
Playback control

Figure 2. The debriefing tool while in use during one of the recorded sessions.

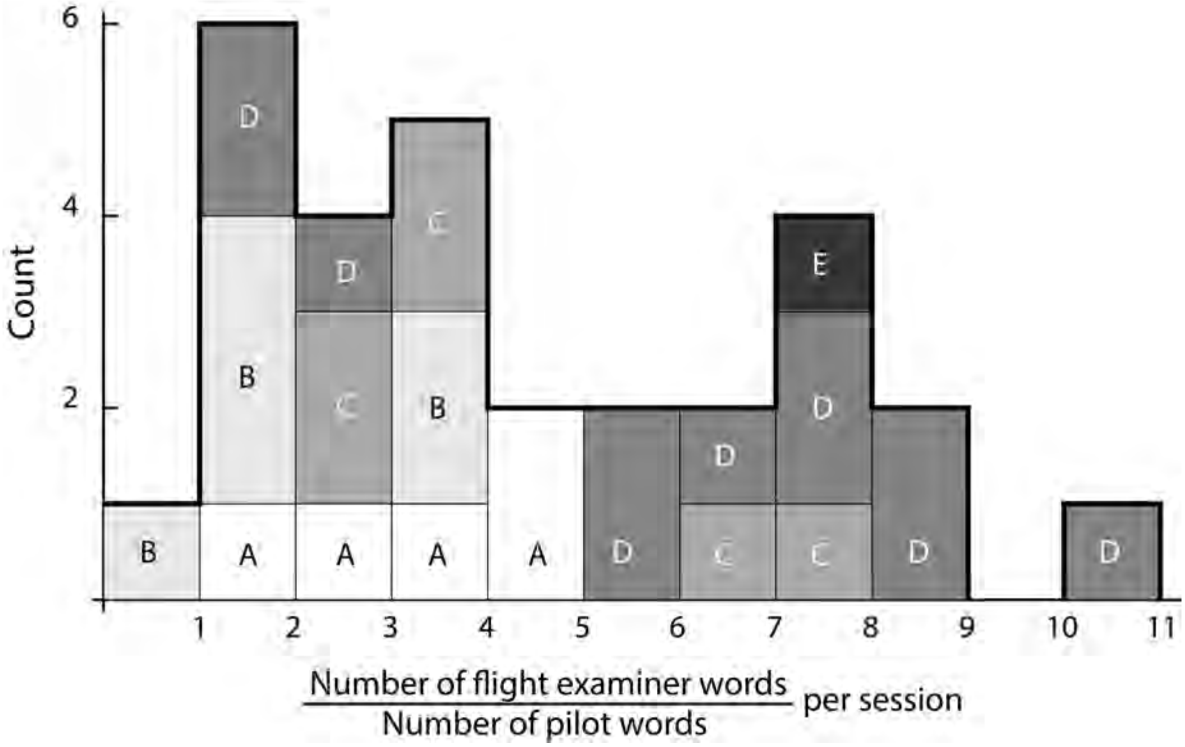

Figure 3. Ratio of number of flight examiner words to number of pilot words per session in five airlines.

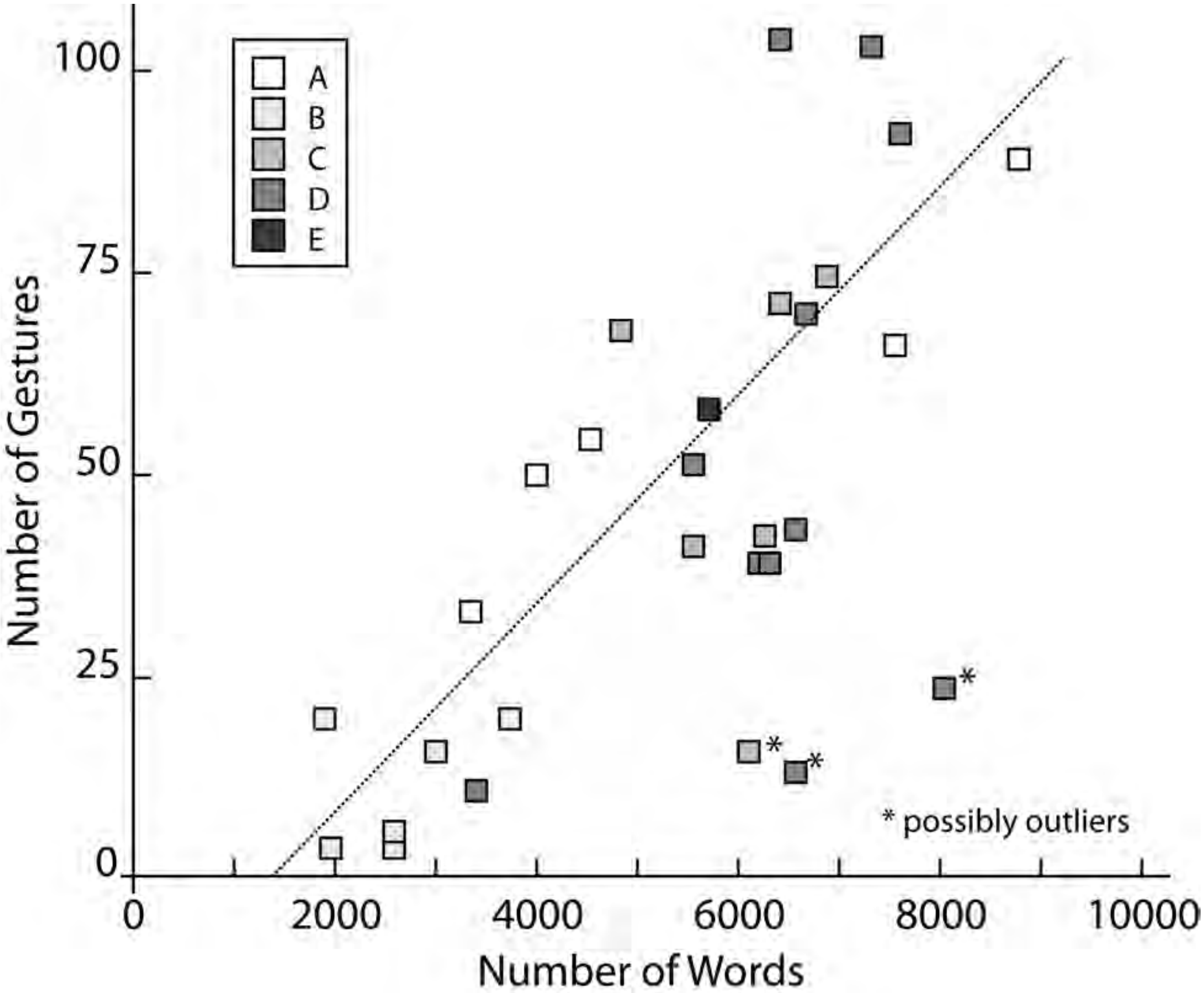

Figure 4. Regression of number of gestures against number of words for debriefing sessions from five airlines (A–E). Asterisk (*) marks three possible outliers (see text).