2007

Memory for Form and Color in Virtual and Real Environments

Matthew Brown Ph.D.

Chris M. Herdman Ph.D.

Jonathan Wade

MEMORY FOR FORM AND COLOR IN VIRTUAL AND REAL ENVIRONMENTS

Matthew Brown, Ph.D., Chris M. Herdman, Ph.D., and Jonathan Wade
Aviation and Cognition Engineering (ACE) Lab
Centre for Advanced Studies in Visualization and Simulation (VSIM)
Carleton University, Ottawa, Canada

Three experiments that differed in terms of environment type (two virtual and one real) required participants to remember the location of 12 unique geometric objects that varied in terms of form (cone, cube, and sphere) and color (blue, green, red, and yellow). The results showed that regardless of environment type (virtual or real), participants showed significantly better encoding/memory for an object's form than for its color. This highlights the importance of emphasizing object shape to assist learning and memory in virtual environments.

Introduction

The recent and impressive technological advancements associated with the implementation of Virtual Environments (VEs) and their widespread use have somewhat overshadowed the observation that humans have difficulty interacting with them relative to real environments (Darken, Allard & Achille, 1998). These difficulties appear whether the VE is completely passive or highly interactive (but see Farrell, Arnold, Pettifer, Adams, Graham, & MacManamon, 2003). Given that VEs are relatively new creations, there is a limited understanding of why humans have difficulty interacting with them relative to real environments.

Understanding the strengths and limitations of how humans encode, store and retrieve object color and form information appears to be a promising candidate for study in this area, especially given the claim that this process is accomplished differently in real and virtual environments. Specifically, object recognition is traditionally argued to provide the basis for representation in visual memory (e.g., Biederman, 1987). Although color information is involved in the segmentation of an object from the background (e.g., Gegenfurtner, Wichmann, & Sharpe, 1998; Jacobs, 1981), the received view is that it does not play a critical role in the storage of object-based information in visual memory (see Oliva & Schyns, 2000; Wichmann, Sharpe, & Gegenfurtner, 2002). More recently, it has been argued that this view may only apply to the 'unnatural' (e.g., artificial, isolated) object representations used in the studies that fostered this traditional understanding. Indeed, it has been shown that color is as important as form in object recognition in more 'natural' (real) environments (Spence, Wong, Rusan, & Rastegar, 2006). This finding is consistent with Steeves, Humphrey, Culham, Menon, Milner, & Goodale's, 2004 claim that scene perception in natural environments is not identical to object perception in artificial (virtual) environments.

The question addressed here is whether VEs and the objects contained therein are mentally processed as being 'natural' or 'unnatural.' If a VE based on a real environment is mentally interpreted as a natural scene, then an object's color should be represented in visual memory to the same degree as its form. If, however, a VE is treated as unnatural in the sense that it is processed as a series of artificial and isolated objects, then an object' form should be more important than its color for remembering information related to that object.

Experiment 1

The question discussed above was operationalized by using computer software to create a VE based on a real environment. The real environment on which this simulated VE was based was a 12 by 12 foot room. The room was populated with 12 objects [one of three geometric primitives (cone, cube, sphere) painted one of four colors (blue, green, red, yellow)] that were randomly assigned (but held constant across participants) to specific locations around the room's perimeter (see Figure 1). These objects were selected because they independently varied along the two dimensions of interest (color and form), thus the ability to remember information specific to each dimension could be analyzed.

The task was to remember the location of each object. Participant's memory was assessed by having them draw an object in its appropriate location on a diagram showing an overhead view of the room. A trial consisted of a 45 second exploration phase followed by a memory test. Participants had to remember the location of all 12 objects on two consecutive trials to complete the experiment.

Method

Participants. Twenty adults, most of which were Carleton University undergraduate students, participated and received either $10 or 0.5% course credit. Participants had normal color vision and were assumed to have normal or corrected-to-normal visual acuity.

Materials. The real environment consisted of a temporary 12 by 12 foot room (henceforth the experimental room). The walls of the experimental room were black opaque plastic sheets that were attached to the ceiling and floor of the encompassing room. The black opaque sheeting was purposefully selected so as to prevent any external cues (e.g., external light) from being visible from within the experimental room. The four experimental room walls were labeled with the numbers "1", "2", "3" and "4." The purpose of doing so was to provide reference points for the subsequent memory test given that no cues (other than the test objects themselves) differentiated one wall from another. The objects were Styrofoam forms (approximately 6 x 6 x 10 inches in size) placed on black pedestals that were approximately four feet in height and evenly spaced around the room's perimeter.

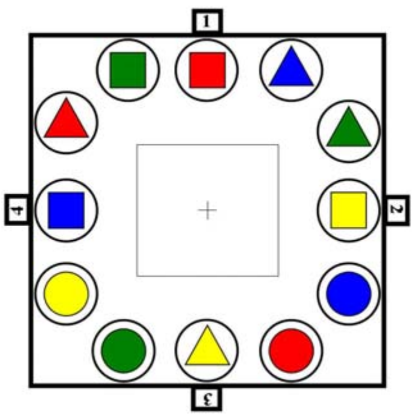

Figure 1

The VE was created using Valve Software's Source engine (which underpins the Half Life 2 gaming platform). The textures (e.g., wall coverings) were created using Adobe Photoshop. This fully synthetic VE was designed to be as visually similar as possible to the experimental room, taking into account relative sizing, textures and lighting. Movement was restricted to an imaginary 5 foot by 5 foot centrally located square. The VE was viewed on a 17 inch color CRT monitor at a resolution of 1024 x 768 pixels slaved to a PC-compatible computer with a dual Xeon processors, two gigabytes of RAM and an ATI Radeon 9800 graphics card.

Participants used a standard mouse to navigate within the room by sliding it to mimic a head rotation and left-clicking to move forward and right-clicking to move backwards. Only head rotations were necessary to view all twelve objects, although some participants showed a preference for being closer to objects that they were attempting to encode.

Procedure. Prior to the beginning of the experimental phase, participants familiarized themselves with the experimental room and the interface used to control their movements. This practice phase was limited to 45 seconds so as to provide participants with an expectation for the duration of each experimental trial. The objects were not visible during the practice phase. The experimenter instructed the participants to note the numbering on the four walls. Participants always started an experimental trial centered in the room facing wall number 1. After the 45 second exploration phase, the computer monitor was blanked and participants were asked to draw the 12 objects in their appropriate location on a bird's-eye-view representation of the room (similar to Figure 1, except that the objects were not shown). There was no time limit for this recall task nor were subjects forced to make a response (they could leave a space blank if they chose to do so).

Results

Two 2 x 2 goodness of fit chi-squares were used to determine whether an object's form or color was more easily encoded and subsequently remembered. One chi-square was used to assess whether an object's form was recalled better than chance, and a second chi-square was used to assess whether an object's color was recalled better than chance. One way to determine whether form or color was more easily recalled is to examine the types of errors participants made. That is, were form errors (e.g., incorrectly guessing 'blue cube' instead of 'blue sphere') or color errors (e.g., incorrectly guessing 'blue cube' instead of 'green cube') more likely to occur? Only instances where an attempted error was made (i.e., where participants made an incorrect response, as opposed to making no response whatsoever) were included in this analysis.

Participants made 533 attempted errors. In order to determine the number of times an object's form or color could be correctly guessed by chance alone, the total number of attempted errors was divided by the number of possible alternatives for either form (3) or color (4). For example, the number of times that an object's color could be correctly guessed by chance alone for the 533 attempted errors is 133 (533/4). This value is seen in the 'correct-expected' cell for 'color' for Experiment 1 shown in Table 1. The number of times that an object's color could be incorrectly guessed by chance alone is simply the difference between the number of attempted errors (533) and the value in the 'correct-expected' cell (133). Thus, the value appearing in the 'incorrect-expected' cell for 'color' for Experiment 1 is 400 (533 – 133). These expected frequencies were then compared to the observed frequencies of correct and incorrect responses, which formed the four cells of both 2 x 2 chi-squares.

Table 1

Experiment			correct	incorrect
1: Virtual Simulated	Color	observed	146	387
		expected	133	400
	Form	observed	277	256
		expected	178	355
2: Virtual Image-Based	Color	observed	125	394
		expected	130	389
	Form	observed	270	249
		expected	173	346
3: Real	Color	observed	125	421
		expected	137	409
	Form	observed	311	235
		expected	182	364

The difference between the likelihood of making a color error compared to a form error is striking. The likelihood of making a color error was at chance, $\chi^2(1) = 1.63$, $p > .20$. In stark contrast, the likelihood of making a form error was significantly less than chance, $\chi^2(1) = 83.27$, $p < .001$.

Discussion

The finding that participants are more likely to make color errors than form errors when an object's location is unknown suggests that participants are preferentially encoding/retrieving object form information than object color information. This is consistent with the view that this VE and the objects contained therein are mentally processed as 'unnatural.' If this VE had been processed as 'natural' (real), then color information would have been as equally as important as form information. If this were the case, then the two chi-square values would not have been as strikingly different.

An alternative account is that the limitations of the software used to generate the VE used in Experiment 1 resulted in it being of sufficiently low visual fidelity that the environment appeared artificial, even though it was based on a real environment. That is, the VE simply looked "fake" and was therefore processed as being 'unnatural.' In order to address this issue, a second VE was created (based on the same real environment), except that panoramic images were used instead of it being fully simulated. This approach was based on the assumption that a VE created using high-quality panoramic images would have a higher visual fidelity than a fully simulated VE.

Experiment 2

The purpose of Experiment 2 was to determine whether a VE with higher visual fidelity than the fully simulated VE used in Experiment 1 would result in participants processing it as 'natural.' The high visual fidelity VE used in Experiment 2 was created by capturing six images with a custom digital camera with six lenses (five lenses pointing outwards and one lens pointing upwards) and then stitching them together using specialized software created by Canada's National Research Council (NRC) to eliminate overlap as well as other visual anomalies. The end result of this stitching process was a panoramic image of the experimental room as it would appear when standing at a given location. If the higher visual fidelity of the VE used in Experiment 2 is sufficient to result in it being processed as 'natural,' then an object's color should be as important as its form and the two chi-squares should not yield strikingly different results.

Method

Participants Twenty Carleton University undergraduate students participated and received either $10 or 0.5% course credit. Participants had normal color vision and were assumed to have normal or corrected-to-normal visual acuity.

Materials Thirty-six panoramic images were used to create the VE. Images were captured at each intersection point of a grid with 1 x 1 foot spacings overlayed on a 5 by 5 foot centrally located square. A custom viewer created by the NRC using Open GL provided the interface that allowed participants to "move" from one panoramic image to the next. Transitions between images were reasonably fluid inasmuch as participants did not state that moving between them interfered with their task.

Procedure The procedure for Experiment 2 was identical to that of Experiment 1.

Results

Participants made 519 attempted errors in Experiment 2. These errors were submitted to the same 2 x 2 chi-square format used in Experiment 1. These data are seen in the 'Experiment 2' row of Table 1. The same pattern of data seen in Experiment 1 is also observed here. That is, the likelihood of making a color error was at chance, $\chi^2(1) = 0.23$, $p > .20$, whereas the likelihood of making a form error was significantly less than chance, $\chi^2(1) = 81.58$, $p < .001$. Using an image-based VE in Experiment 2 instead of a fully simulated VE did not affect the relative probability of making color and form errors.

Discussion

Increasing the visual fidelity of the VE did not result in it being mentally processed as 'natural.' Participants continued to make significantly more color errors than form errors when an object's location was unknown, which suggests that they are preferentially encoding/retrieving object form information. This pattern of data is consistent with the view that this image-based VE is treated as 'unnatural.'

An explanation for the inability to create a VE that is mentally processed as 'natural' is that the defining characteristic of 'naturalness' is not the realism of the environment itself, but rather the realism of the objects contained therein. That is, although the VEs used in Experiments 1 and 2 are based on a real environment, the objects used in these experiments are, to some degree, artificial. The colored geometric primitives used here are fundamentally different than natural objects (e.g., a telephone) in the sense that they are devoid of semantic meaning. Perhaps it is this lack of meaning that resulted in them being treated as 'unnatural' despite the fact that they were presented in a high visual fidelity VE that was based on a real environment. In order to test this hypothesis, memory for object location in the real environment (on which both VEs were based) was assessed using these same objects.

Experiment 3

Experiment 3 sought to address the question of whether the artificiality of the objects used in Experiments 1 and 2 caused participants to treat these environments as 'unnatural.' Participants' memory for the same objects and their respective locations was assessed in a real environment. If the artificiality of the objects themselves determines how a scene is mentally processed (i.e., as natural or unnatural), then participants should show preferential encoding/memory for form than for color, despite this information being learned in a real environment.

Method

Participants. Twenty Carleton University undergraduate students participated and received either \$10 or 0.5% course credit. Participants had normal color vision and were assumed to have normal or corrected-to-normal visual acuity.

Materials. The real environment used in Experiment 3 is described as the, "experimental room" in the Method section for Experiment 1.

Procedure. The procedure for Experiment 3 was identical to that of Experiments 1 and 2, except for the following differences. The practice session consisted of having participants explore the real environment instead of viewing it on a computer monitor. During this time, the objects were covered such that neither their form nor their color was visible. The experimenter noted a 5 by 5 foot area demarcated on the floor and instructed participants to limit their movement to this square. The purpose of limiting participant movement was to mimic the limited number of viewpoints available in Experiments 1 and 2. Once the participants had familiarized themselves with the experimental room and understood their task, they were asked to leave the experimental room at which point the experimenter removed the object covers. The experimenter then exited the experimental room and asked the participant to enter with their gaze directed towards the floor such that they could not see the objects before the first experimental trial started. Participants situated themselves in the centre of the room (demarcated on the floor by a '+' sign) facing wall number 1. During this time, the experimenter closed the gap in the wall through which the participant entered the experimental room so that each wall appeared to be visually identical save for the numbered signs. The experimenter verbally instructed the participant to 'start' at which point the participant began viewing the objects. After 45 seconds had elapsed on a stopwatch, participants were instructed to 'stop' and to direct their gaze downwards (so as not to view the objects) as they exited the experimental room.

Results

Participants made 546 attempted errors in Experiment 3. These errors were submitted to the same 2 x 2 chi-square format used in Experiments 1 and 2. These data are seen in the 'Experiment 3' row of Table 1. The same pattern of data seen in Experiments 1 and 2 is also observed here. That is, the likelihood of making a color error was at chance, $\chi^2(1) = 1.29$, $p > .20$, whereas the likelihood of making a form error was significantly less than chance, $\chi^2(1) = 137.12$, $p < .001$.

Discussion

The finding that form information was preferentially encoded/retrieved in the real environment used in Experiment 3 suggests that it was mentally interpreted as being ‘unnatural.’ This is consistent with the view that the realism of the objects themselves in an environment (whether it be virtual or real) is critical in determining whether that environment is mentally processed as ‘natural’ or ‘unnatural.’ Evidently, the objects used in this series of experiments were perceived as being sufficiently artificial that it resulted in both the virtual and real environments being mentally processed as ‘unnatural.’ If the environments used here had been mentally interpreted as ‘natural,’ then color information should have been as critical as form information in terms of acquiring knowledge about object location. If this were the case, then this pronounced difference between the frequency of form and color errors would not have been observed. This assertion is based on Spence et al.’s (2006) finding that color information, by default, plays a critical role in the encoding and recollection of images of natural environments. An alternative explanation for the unexpected finding in Experiment 3 is that the task demands of the experiment itself (i.e., testing memory for object location), may have forced participants to actively encode objects as being discrete and isolated from their surrounding environment instead of passively allowing their visual system to use color and other surface cues to segment the objects from the background.

Conclusions

The results of the present research clearly and consistently showed that participants encoded object form before object color, regardless of whether the objects were presented in a VE (with different levels of visual fidelity) or in a real environment. This finding is consistent with the view that color information has little impact on visual memory (Oliva & Schyns, 2000; Wichmann et al., 2002). It is acknowledged, however, that color information must have been encoded at some point given that this feature was the only way to visually differentiate between objects of the same form (e.g., the blue and red cubes). That said, it would appear that an object’s color is typically encoded after its form, and perhaps only when it is necessary to do so. These findings support the view that an object’s form is critically important in the quick and effective encoding/retrieval of its location. Consequently, over-emphasizing an object’s form (by artificially enhancing it) may prove to be an easy and effective method of facilitating learning in VEs.

References

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 9, 115-147.

Darken, P. D., Allard, T., & Achille, L. B. (1998). Spatial orientation and wayfinding in large-scale virtual spaces: An introduction. *Presence*, 7, 101-107.

Farrell, M. J., Arnold, P., Pettifer, S., Adams, J., Graham, T., & MacManamon, M. (2003). Transfer of route learning from virtual to real environments. *Journal of Experimental Psychology: Applied*, 9, 219-227.

Gegenfurtner, K. R., Wichmann, F. A., & Sharpe, L. T. (1998). The contribution of color to visual memory in X-chromosome-linked dichromats. *Vision Research*, 38, 1041-1045.

Jacobs, G. H. (1981). Comparative color vision. New York: Academic Press.

Oliva, A. & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41, 176-210.

Spence, I., Wong, P., Rusan, M., & Rastegar, N. (2006). How color enhances visual memory for natural scenes. *Psychological Science*, 17, 1-6.

Steeves, J. K. E., Humphrey, G. K., Culham, J. C., Menon, R. S., Milner, A. D., & Goodale, M. A. (2004). Behavioral and neuroimaging evidence fro a contribution of color and texture information to scene classification in a patient with visual form agnosia. *Journal of Cognitive Neuroscience*, 16, 955-965.

Wichmann, F. A., Sharpe, L. T., & Gegenfurtner, K. R. (2002). The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 509-520.

Acknowledgements

This research was supported by a Natural Sciences and Engineering Research Council (NSERC) strategic grant. Correspondence should be directed to Dr. Chris M. Herdman, Scientific Director, Centre for Advanced Studies in Visualization and Simulation, Dept. of Psychology, Carleton University, Ottawa, ON. Canada, K1S 5B6. Email correspondence: chris_herdman@carleton.ca

PILOT SELF-REPORTS AS A RELIABLE MEANS OF COLLECTING SAFETY RELATED OPERATIONAL DATA

Sue Burdekin
University of New South Wales at the Australian Defence Force Academy
Canberra, Australia

Self-reporting is widely used throughout the world in various situations ranging from medical self-diagnosis through to the workplace and educational application. However, self-reporting as a means of gathering operational data has not been accepted as a reliable tool in the aviation industry. Much has been written on self-reporting in other industries with varied results. However, job type has been highlighted as a moderator (Harris and Schaubroeck, 1988; Burdekin, 2003), and Mabe and West (1982) proposed that if individuals understand the dimension in question, accept the dimension, and perceive that the assessment will not be used against them, self-assessments would be more accurate. A study utilizing Australian military pilots in an F/A-18 simulator found that pilots were able to recall and self-report their own behaviours after the 'flight' and that their self-reports correlated with ratings of the same behaviours made by an independent observer (Burdekin, 2003). The present civil pilot self-report study was conducted at a low cost airline in Europe during normal revenue raising flights. The crossed design included captain/observer, first officer/observer and captain/first officer self-reports. Significant correlation between raters was reported. Some issues were revealed concerning the first officers' data and behavioural markers. These are discussed along with future application for operational self-reporting in the aviation industry.

Introduction

Self-assessment is used widely for a variety of applications including patient self-assessment and monitoring of minor ailments in medicine; self-assessment of vital signs in sport and exercise environments; employee self-assessment in an industrial or administrative workplace; and, student self-assessment in an educational setting. When self-assessment is a formal requirement a self-report will be generated.

Much has been written concerning the reliability of self-reporting and its contribution in the workplace. In some situations the concept has been accepted as standard practice, for example, workplace appraisal. Yet in other situations self-reporting has been less likely to be adopted, for example, in-flight operations in the aviation industry.

The literature generally supports three characteristics of accurate self-evaluation – relatively higher intelligence, high achievement status and internal locus of control (Mabe & West, 1982). Harris and Schaubroeck (1988) reported that job type was a moderating factor in self-peer and self-supervisor ratings. They suggested that skilled, well-trained workers performing interdependent tasks that have been clearly defined were more capable of accurately self-reporting their behaviour. Furthermore, it has been proposed that if individuals understand the dimension in question, accept the dimension, and perceive that the assessment will not be used against them, self-assessments will be more accurate (Mabe & West, 1982).

Very little has been written on operational self-reporting in an aviation environment. Burdekin (2003) conducted a study in Australia with 30 military F/A-18 pilots flying a predetermined (but unknown to the subject) mission in the simulator. The results indicated that pilots were able to recall and self-report their own behaviours after the 'flight' and that their self-reports correlated with ratings of the same behaviours made by an independent observer. Military pilots are highly skilled, comprehensively trained and combat ready. The level of commitment and behaviour required of these pilots is well defined and because of this they were able to make informed appraisal of their own actions and reactions according to the operational circumstances they encountered during the flight.

In the military pilots' study it was hypothesized that because of egocentric bias, which refers to the "underlying premise that self-report ratings are fundamentally biased whilst other objective raters share common rating perceptions" (Burdekin, 2003 p.24), pilots would over-estimate their performance. However, results indicated that the pilot self-report ratings were, in fact, marginally lower than the observer ratings, highlighting the fact that the pilots were more critical of their own performance. This might indicate that pilots have the training and ability to be more diagnostic concerning their own behaviour.

The aim of the present study was to determine if civil airline pilots operating normal short haul revenue raising flights were able to elicit the same behavioural ratings as those of an independent, objective observer. Captains and First Officers were asked to not only assess their own behaviour but also to assess the behaviour of each other using the same behavioural criteria.

Method

Participants

The participants in this study were all Air Transport Pilot Licence (ATPL) rated, employed by easyJet and based at Geneva. The pilots flew as a crew consisting of one captain and one first officer. As expected in any operational environment the age and experience of the pilots varied, but all were appropriately endorsed on the aircraft and held the ratings essential for employment as an airline pilot. Every participant in this study was a volunteer.

Two observers were trained for the study. One was a psychologist involved with the design of the study, who had aviation qualifications to (frozen) ATPL level, and the other was a company line pilot endorsed on the aircraft type who volunteered for the task. Both observers graduated from a training course in the Line Operations Assessment System (LOAS) at the Airbus Training Centre in Toulouse.

Design

As is the case for most applied research it was difficult to control for all variables. However, in an attempt to minimize the external influences on the study, the Geneva base of the easyJet airline operation was chosen because of the fact that it only operated A319 aircraft and most of the pilots were Swiss nationals. Additionally, at the time of the study, the route structure of easySwiss (as it is known internally) was limited to eight destinations.

The rating protocols were designed by the chief researcher and two subject matter experts (SMEs) from the Head Office Safety Department at easyJet Luton, who consisted of an operational captain and a psychologist. The protocols were influenced by the Australian Defence Force Mission Operations Safety Audits (MOSA) and the LOAS methodology from Airbus.

Company members selected eight categories of crew behaviour that represented areas of special interest to them: briefing, contingency management, monitor/cross check, workload management, situational awareness, automation management, communication, and problem solving/decision making. Each behavioural category was given a comprehensive descriptor. For example, the descriptor for the category ‘briefing’ was:

The required briefing was interactive and operationally thorough. Concise, not rushed and met SOP requirements. Bottom lines were established. Roles and responsibilities were defined for normal and non-normal situations. Workload assignments were communicated and acknowledged.

In order for the pilots to make a more informed rating choice, a series of specific ‘word pictures’ were given to each category of behaviour, ranging from a grading of 1 to 5 (refer Table 1).

Table 1. Grading/Word Pictures for the Behavioural Category ‘Briefing’

1.	Unsatisfactory briefing standard. Briefing duration and crew interaction minimal. Available company resources not utilized to a satisfactory standard. SOPs not adhered to.
2.	Basic briefing conducted with limited crew interaction. Incomplete use of available resources and workload allocation limited. SOP briefing structure loosely adhered to.
3.	Crew operates in accordance with SOP briefing structure. Interactive briefing conducted in a timely manner, utilizing available resources to an adequate standard.
4.	Effective crew briefing conducted utilizing all company/non-company information. Proficient time and workload management with clear interaction and allocation of duties amongst crew.
5.	Comprehensive and operationally thorough briefing conducted to a high standard. Excellent crew interaction, participation and understanding. All available briefing resources utilized and clear and concise workload allocation amongst crew.

If a behaviour was rated as particularly good, or not up to standard, pilots were asked to identify the phase of flight in which this occurred in (pre-departure/taxi; take-off/climb; cruise; descent/approach/land; taxi-in; or all). In order to simplify the answer sheet and with time constraints in mind, pilots were given neutral operational 'key words' or phrases that they could highlight in a positive or negative sense to provide a further explanation – for example, interactive, conformity with SOP structure, timely, workload, communication etc. If the pilot wished to expand on this in his own words there was ample space to do so.

The company was also interested in the interaction between its crews and air traffic control, ground support, aerodromes and passengers. Space was allocated on each flight sector form for open comments concerning these topics.

Procedure

Data were collected over the course of a pilot's working day. There were two shifts that could be flown. The first commenced with sign-on at 5:00am and concluded around 2pm, depending on the routes flown and whether or not there were any delays. The second shift commenced at around 1pm and finished at approximately 10pm. The working day consisted of four flight sectors. All flights originated at home base Geneva and each pair of sectors terminated back at Geneva. For example, the first flight (therefore the first sector) might be from Geneva to Paris, Orly airport, and so the second sector would be back to Geneva from Orly. All crewmembers and the observer traveled to the aircraft together and remained together for the entire shift.

In an attempt to minimize the impact of this study on the crews and the airline operations in general, all the instructions for the pilots were written on a self-briefing form and given to the participant prior to the pre-flight briefing. The period of data collection started during the pre-flight briefing and finished when the crews signed off for the day. Because of the nature of collecting data in the field a degree of flexibility had to be allowed. Some pilots chose to write up their self-reports in the aircraft, and others completed the forms when they had returned to the office after the shift.

For a more comprehensive analysis by the company's Safety Department each pilot participant was requested to fill out a form outlining demographic details including: an age range, hours on type, total hours, total years flying, years experience on advanced automated types – for example, FMC with VNAV/LNAV, approximate time since last CRM course, and details of any human factors training. The company was also interested in cross-referencing some other roster information with a fatigue study that it was conducting with British crews at the time. The demographic information was collected in terms of approximate figures and was not able to be personally identified. It is mentioned here for the information of readers who may see an application for such information in further experimental studies, or in operational practice. Since the present study is concerned with the ability of the pilots to self-report, the information contained in this paragraph will not be referred to again in this paper.

An inter-rater reliability test was conducted to ensure that the observers' ratings were consistent across the categories of behaviour. Each observer, using the experimental study protocols, independently rated the behaviour of easyJet crews who appeared in three simulator Line Operations Flight Training (LOFT) videos that were made by the company's Safety Department.

Prior to the commencement of the in-flight data collection, the researcher and company management members addressed the pilots union and explained the aim of the experiment. The union did not object to the study, providing that the data were protected and only volunteer pilots were involved. All information collected was anonymous and could not be identified by pilot name, flight number or date. The researcher was the 'gate-keeper' for all raw data.

Results

Sixty flight sectors were observed, but 59 sectors were included in the study due to one first officer subsequently declining to participate after he had initially volunteered.

Having established the inter-rater reliability of the two observers, the total observer (OBS) ratings across all behavioural markers were collapsed and compared with the captain's (CPT) and first officer's (F/O) ratings. The total ratings of all of the captains were also compared with the total ratings of all of the first officers. The results were significant in all cases (refer Table 2) indicating that the pilots' ratings of their own performance agreed with the ratings of the independent observers.

Table 2. Ratings across all behavioural markers

Rater	Mean	sd	N	r
OBS CPT	3.995 4.000	.36345 .58599	432	.344 **
OBS F/O	3.995 3.7394	.36345 .53281	376	.206 **
CPT F/O	4.0000 3.7394	.58599 .53281	376	.169 **

** *correlation is significant .001*

In order to determine how the individual behavioural markers contributed to the overall results, the ratings were compared by behavioural category (refer Table 3). All markers except 'automation management' and 'problem solving/decision making' were found to be significant contributors when comparing the observer's and captain's ratings. Only half the markers were successful when comparing the observer's ratings with the first officer's ratings. In this case the 'automation management' marker was a contributing factor. However, when comparing the captain's ratings with the first officer's ratings only the 'communications' behavioural marker was significantly correlated. This indicates that the captains appear to have accurately interpreted how to rate a pilot's performance utilizing the behavioural marker system, whereas first officers were having trouble with the assessment. Furthermore, the results show that the first officers were more critical of both the captain's and their own performance. It is important to note that First officers did not attempt to over inflate their performance, as some theories predicted.

Insufficient training in the study, lack of experience and inadequate understanding of what was required of them may have influenced the first officer's interpretation of the behavioural categories and how to rate them in-flight. In the present study no company promotion of self-report techniques or the program was undertaken apart from the support that was requested from the pilot's union. It is also possible that the categories of behaviour themselves might need adjustment or clarification, either in the choice of behaviour or the 'word pictures' describing it.

Table 3. Correlations by behavioural category

Behavioural Marker	OBS/ CAPT	OBS/ FO	CAPT/ FO
Briefing	.630 **	.427 **	.275
Contingency Mgt	.361 **	.056	.000
Monitor-x-check	.278 *	-.098	-.043
Workload Mgt	.364 **	.164	.132
Situational Awareness	.271 *	.368 *	.115
Automation Mgt	.109	.301 *	.195
Communications	.457 **	.441 **	.335 *
Problemsolve/Decision Making	.265	-.232	-.116

* *correlation is significant .005*
** *correlation is significant .001*

No statistically significant correlation could be found with the pilots' choice of key words to provide a further explanation of the rating selection. However there were some comments offered in long hand text if a rating was particularly good or conversely particularly poor, indicating that pilots were prepared to expand on their choice of rating selection.

Conclusion

The primary aim of the present study was to determine whether civil airline pilots operating normal, short haul, revenue raising flights were able to generate the same behavioural ratings of their performance as those of an independent; objective observer. The results indicated that airline pilots are able to reliably self-report on the performance of themselves, and another technical crewmember. These results lend further support to the previous findings that job-type appears to be an influential factor for self-report reliability (Harris and Schaubroeck, 1988; Burdekin, 2003). Deeper analysis of the results indicates that first officers might need more training in how to make a behavioural judgment using the rating scale adopted.

Civil airline pilot self-assessment of operational behavioural issues has the potential to reduce costs and provide insight into in-flight operations from front-line specialists. One application of pilot self-reporting is in the field of flight data monitoring (FDM) supported by Reisinger et al (2005). They believe that the FDM system could be more efficient if it could be matched to a formal system of operational crew self-reports explaining the details of the flight. They propose that flight crews should conduct their own flight data analysis by being permitted access to the FDM information from their own flights, thereby enabling them to add value to the flight data, such as recognizing potential threats and predicting when error could occur, along with their personal coping methods and management strategies. The objective would be to gain a greater insight into why certain decisions were made in flight. If pilots are made aware of this type of information, it may be utilized in identifying and enhancing individual pilot training.

References

Ashford, S. J. (1989) Self-Assessments in Organisations: A literature review and integrative model. In Research in Organisational Behavior, Vol. 11, 133-174. JAL Press Inc.

Burdekin, S. G. (2003) Mission Operations Safety Audits: Measurements of behavioural performance or non technical skills from military air crew. Spotlight 0403. Directorate of Flying Safety. Canberra.

Carver, C. S., & Scheier, M. F., (1981) Attention and Self-Regulation: A control-theory approach to human behavior. Springer New York..

Drewes, G. & Runde, B. (2002) Performance Appraisal. In Psychological Management of Individual Performance. Editor Sonnentage, S. John Wiley & Sons Ltd.

Fletcher, C. (2002) Appraisal: An individual psychological perspective. In Psychological Management of Individual Performance. Editor Sonnentage, S. John Wiley & Sons. Ltd.

Harris, M. M., & Schaubroeck, J. (1988) A Meta-Analysis of Self-Supervisor, Self-Peer, and Peer-Supervisor Ratings. In Personnel Psychology Inc. Vol. 41, 43-62

Mabe, P. A. & West. S. G. (1982) Validity of Self Evaluation of Ability: A review and meta-analysis. In Journal of Applied Psychology. Vol 67. No. 3. 280-296

Reisinger, D., Sporer, S., & Knoll, G. (2005) Flight Data Analysis – A New Approach. In proceedings from the International Society of Air Safety Investigators Conference. September 2005. Fort Worth.

Acknowledgement

The author would like to thank the administrative and operational staff of easyJet Geneva for their cooperation and assistance in collecting the data for this research with special thanks to Captain Philippe Pilloud. Thanks must also be given to the staff of the easyJet Safety Department at Luton and in particular Captain Simon Stewart and Rafeef Abboud for their contribution to the design of the experimental protocols. The author would like to acknowledge the very kind assistance of Captain John Scully from Airbus and Captain Peter Griffiths from easyJet.