Development of a Taxonomy of Causal Contributors for Use with ASAP Reporting Systems

FAA Grant # 99-G-048

Technical Report

June 2005

Prepared by:

Kelley J. Krokos, Ph.D.
David P. Baker, Ph.D.
American Institutes for Research
1000 Thomas Jefferson Street NW
Washington, DC 20007-3935

Prepared for:

Federal Aviation Administration

Eleana Edens, Ph.D.
Human Factors Research and Engineering Division
Flight Deck/Aviation Maintenance
Room 907A
800 Independence Avenue, SW
Washington, DC 20591

Thomas Longridge, Ph.D.
Flight Standards Service
Voluntary Safety Programs Office
45005 Aviation Drive, Suite 203A
Dulles, VA 20166

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

ACKNOWLEDGEMENTS

The American Institutes for Research (AIR) recognizes the significant contribution of several organizations. First, AIR recognizes the Federal Aviation Administration (FAA), and in particular, Eleana Edens at the Flight Deck/Aviation Maintenance Branch at the Human Factors Research and Engineering Division (AAR-100) and Tom Longridge and Doug Farrow at the Voluntary Safety Programs Branch at the Flight Standards Service (AFS-230). In addition to providing funding for the development of the taxonomy, they also provided guidance and support along the way.

AIR also recognizes its research partner, American Eagle, Inc. (EGL). Safety Office officials including Captain Matt Wise, Captain Craig Chilcote, and Vickie Toman were instrumental in the development of the taxonomy in several ways. First, they provided information, guidance, and advice regarding the ASAP program and its operation, both in general and at EGL. Second, they provided organizational support and encouragement to EGL SMEs to participate in developmental activities including sorting tasks, web-based surveys, and usability tests. Third, they provided numerous hours of participation in developmental exercises, consultation, and feedback regarding the taxonomy. Finally, they believed in the project during its development and continue to support its use and development now that it has been implemented into their new online reporting system. In addition to Safety Office employees, AIR also thanks the pilots at EGL who contributed their time and resources while participating in developmental activities.

Finally, AIR would like to recognize the role that the FAA and the Air Line Pilots Association, Int'l played in the development process. Representatives from both organizations participated in tests of the usability and reliability of the taxonomy. As experts in aviation performance and safety, their contributions were also invaluable.

INTRODUCTION

It is generally recognized that humans have a finite set of cognitive and physical resources and that when faced with complex target rich environments, can and will commit errors (Reason, 1990; IOM Report, 1999). Aviation errors can be particularly costly, both in terms of human life and loss of resources. Consequently, egregious or intentional human errors, especially those resulting in loss of life are generally investigated with the intent of identifying fault. However, typical day-to-day performance of aviators not resulting in an accident can also reveal a great deal of information regarding human performance, which can be used to understand underlying causes of errors. The Federal Aviation Administration's (FAA) Aviation Safety Action Program (ASAP) was designed for this non-punitive investigative purpose. The focus is on discovery of the causal contributors to incidents, and the development of solutions to reduce or eliminate those contributors. ASAP programs have enjoyed significant success; a substantial number of airlines have developed and implemented an ASAP reporting system.

A commonality among voluntary ASAP reporting systems across airlines is that they typically require a crewmember to submit general information regarding the nature of the incident, as well as specific information regarding the perceived cause. For example, ASAP incident reporting forms typically require the crewmember to describe the pilot submitting the report (e.g., seat position, flying time) and the flight conditions (e.g., weather, phase of flight) that immediately preceded the event. Space is also provided for a short text narrative that describes the event, the causal contributors that precipitated it, and suggestions for preventing its reoccurrence.

The qualitative data that results from these text narratives is rich in information. However, in order to allocate resources effectively, organizations must identify patterns or trends in the data. This creates unique challenges, because organizational analysts must read the text-based narratives individually and classify or sort the causal factors into groups. This type of analysis is difficult and time consuming. In addition, this type of analysis does not generally allow for comparison of data across different safety programs (e.g., FOQA and ASAP data), or across organizations within an industry (e.g., American Airlines® and United Airlines). Finally, this sort of analysis is generally conducted by an analyst, who despite being internal to the organization, may or may not be a pilot. In addition, the analyst is in essence a third party contributor in that he or she was usually not present when the incident occurred; the analyst's classification may be made long after the incident actually occurred and on the basis of a short narrative that contains little detail.

In sum, ASAP has great possibilities but has yet to be used to its fullest potential. In response to a request from the FAA's Office of the Chief Scientific and Technical Advisor for Human Factors (AAR-100), AIR reviewed the scientific and technical challenges associated with quantifying human factors issues in ASAP reports. Based on that review, we recommended developing a taxonomy for quantifying issues that are documented within ASAP. Specifically, our project had three primary goals. First, we sought to develop a comprehensive ASAP taxonomy that pilots would use to classify the incidents they report. Such a taxonomy would provide carriers more accurate, summative information on the issues that their crews face first-hand during typical line operations. Second, we sought to evaluate to effectiveness of the taxonomy by applying it to actual ASAP reports and statistically assessing its reliability. Such tests are critical in determining a taxonomy's true utility. Finally, we sought to embed this taxonomy within a searchable data collection and reporting tool. Doing so will streamline the process of collecting, managing, and reporting ASAP data. As a result, carriers' limited resources can be devoted to more goal-directed tasks such as problem identification, analysis, and resolution.

This technical report details the development of the taxonomy. Section one, Background, details the literature review and planning stage of the taxonomy development. Section two, Taxonomy Development, describes two phases of development and one phase of testing. The Discussion presented in Section three summarizes what is currently known about the taxonomy and its use. Finally, the Next Steps section proposes ideas for future research.

BACKGROUND

Literature Review

The first step in the taxonomy development process was to conduct a thorough review of the available literature regarding existing taxonomies of human error, accident/incident reporting systems, and data collection tools (Beaubien & Baker, 2002a, 2002b). The review suggested that an ideal ASAP reporting form should collect five major categories of information. These include: crewmember and flight demographic information (e.g., seat position, flying experience, flight number, origin and destination, etc.), antecedent conditions (e.g., weather, meteorological conditions, air traffic, etc.), human factors (e.g., crew processes, taskwork behaviors, physiological limitations, etc.), consequences and outcomes (e.g., loss of control, runway incursion, fire, etc.), and lessons learned (e.g., suggestions for preventing similar occurrences, an assessment of the incident's safety implications, etc.).

During our review, it became clear that no existing taxonomy, reporting system, or data analysis tool met all these important information needs. Several taxonomies, such as the Human Factors Analysis and Classification System (HFACS) and Line/LOS Checklist (LLC), initially seemed promising. However, a careful analysis revealed that these taxonomies were too coarse to identify specific operational problems or to suggest remedies for those problems. Likewise, several data collection tools, such as the University of Texas Demonstration ASAP Incident Reporting Form, also seemed promising. Unfortunately, a careful analysis revealed that the Incident Reporting Form focused more on demographic and environmental factors than on substantive human factors issues. Lastly, our review also revealed several systems (e.g., JANUS, Systematic Incident Analysis Model [SIAM]) that initially seemed promising, but were largely undocumented at the time of our review, making it difficult to determine how their data were collected, organized, or analyzed (Pounds & Isaac, 2002; Australian Transportation Safety Bureau, 2000).

Our review also revealed how a classification system might be evaluated. Fleishman and Mumford (1991) suggest that internal validity, external validity, and utilitarian concerns are important assets and should drive taxonomy development and evaluation. Specifically, internal validity is proposed to be comprised of individual and overall reliability, the existence of a mutually exclusive and exhaustive list of variables, and statistical clustering of the variables. External validity is said to consist of successful cross validation and the ability to fill in knowledge gaps. Finally utility includes the promotion of communication, efficient use of resources, solution to applied problems, and wide acceptance by users. Despite the appeal of these criteria, we were unable to evaluate existing systems on these factors because so little development and analysis information was available.

Based on these deficiencies, the remainder of our review focused on identifying the specifics of how a sound and useful taxonomy could be developed and what it would look like. First, as previously stated, our review suggested that an ideal ASAP reporting form should collect information regarding crewmember and flight demographic information, antecedent conditions, human factors, consequences and outcomes, and lessons learned. We believe that the information in each of the five categories should be organized into one or more lists of exemplars. However, each list should be only one level “deep.” For example, crew processes would be described using a generic list of teamwork behaviors (e.g., communication, coordination, decision-making, and so forth).

To enhance internal and external validity, and utility, we propose that the system should be developed with significant pilot input. Currently, causal contributors to aviation incidents are generally identified by an analyst on the basis of the text narrative contained in the ASAP report. While a technical expert, the analyst may or may not be a pilot. In addition, the analyst was likely not present when the incident occurred and must make causal designations on the basis of what may be a short or vague narrative. We propose to build a system whereby the pilot who submits the report will determine causal contributors and use the taxonomy to categorize them. If pilots are to be the primary users of the taxonomy, then we propose to build the taxonomy using data gathered from pilots. We believe that a combined theoretical and empirical approach that includes guidance from the literature on human error as well as input from technical experts with regard to safety and the ASAP system will yield the most reliable, valid, and usable taxonomy.

Supplemental information should be collected separately. Specifically, information regarding the order of occurrence (e.g., in the chain of events), relevance (i.e., primary vs.

contributory cause), relationship with others (e.g., co-pilot, dispatch, ATC, maintenance), and other relevant factors. For example, the “crew processes” field could be combined with the “relations with others” field to create various combinations, such as “communication with ATC,” “communication with dispatch,” and “communication with flight attendants.”

Finally, to enhance utility, we envision that the entire taxonomy development process must be widely communicated to the industry, with all relevant documentation contained in a centralized repository. Therefore, we propose that any taxonomy should be documented and disseminated in sufficient detail so that potential users can determine if the taxonomy addresses their needs. This dissemination process could involve the development and maintenance of a web site that provides potential users with up-to-date information regarding the taxonomy and its associated validation research (e.g., technical reports, research protocols, instruments, briefing slides, etc.).

Once the literature review was complete, we developed the taxonomy for use with ASAP systems. The goal was to build validity and utility into the system by addressing the evaluation criteria at each stage of development. The steps involved in this process and our findings at each step are described in detail below.

TAXONOMY DEVELOPMENT

Overview

The development of the taxonomy consisted of several phases. Phase I involved identifying potential causal factors and then developing a two-tier structure for the system. First, a list of causal factors was generated by reviewing existing systems. SMEs then engaged in a series of sorting tasks using these factors. Seven high-level causal categories were identified statistically from this sorting data. Once the seven causal categories were identified, Phase II involved verifying that all the causal factors could be reliably sorted into one of the seven categories. SMEs participated in this developmental task via a web-based survey. The results were used to develop a draft version of the taxonomy that consisted of seven high-level causal categories, each containing a number of more specific causal factors. Finally, Phase III investigated the reliability and usability of the draft taxonomy. In this phase, ASAP meeting attendees[1] used the draft taxonomy to classify de-identified ASAP reports. Reliability was calculated as a function of agreement among the raters, and ease of use was calculated as the mean of SME usability ratings. The following section describes the steps in each of the three

[1] There were seven attendees at the first ASAP meeting, and six attendees at the second ASAP meeting.

phases in greater detail.

Phase I

Generate Comprehensive List of Causal Factors

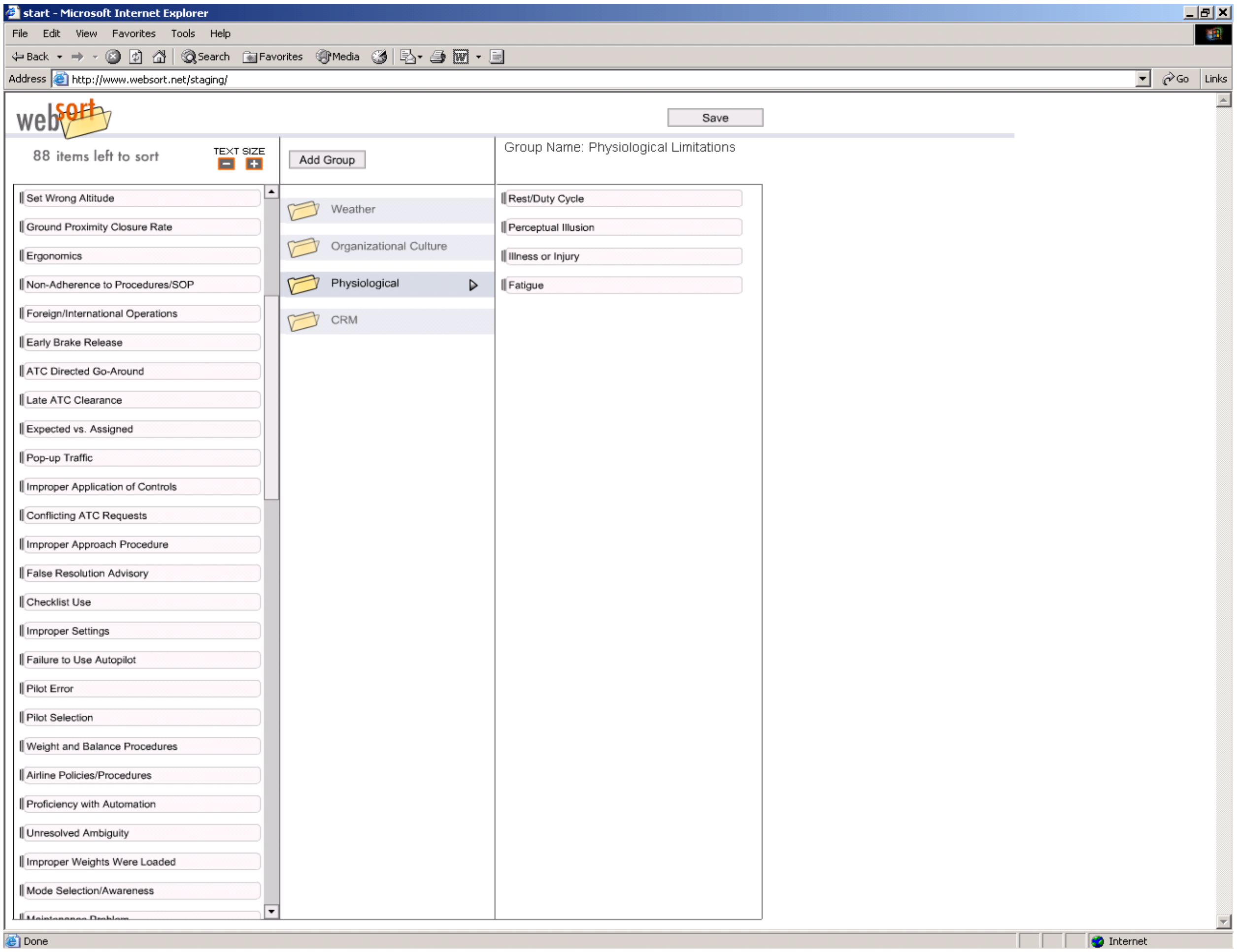
A mutually exclusive and exhaustive list of factors is one contributor to internal validity (Fleishman & Mumford, 1991; see Beaubien & Baker, 2002a, 2002b). To this end, the first step in developing the taxonomy involved generating a comprehensive list of causal factors. De-identified ASRS/ASAP reports, theoretical models of human error, human factors textbooks, and the like were reviewed. Approximately 300 causal contributors consistently appeared in these sources and were included in the list.

Unfortunately, an initial review by SMEs suggested that the list of 300 factors was unwieldy. For example, there were a number of redundancies in the factors (e.g., *ATC Communication Error*, *ATC Miscommunication*, and *ATC Communication too Fast*). Consequently, the list of 300 factors was culled down and simplified by combining redundant factors. For example, *ATC Communication Error*, *ATC Miscommunication*, and *ATC Communication too Fast* were combined to a single factor labeled *Communication/Advocacy* because poor communication was the major cause in all three instances. The result was a more concise list of 94 factors.

Develop Causal Categories

A second contributor to internal validity is statistical clustering of variables (Fleishman & Mumford, 1991). To this end, we sought to group the 94 factors statistically into groups based on how SMEs grouped them. The taxonomy should be only one level “deep,” meaning that there should be one group of high-level causal categories with a single layer of more detailed factors underlying these categories. Consequently, once the list of 94 factors was identified, the next step was to develop a list of high-level categories that could encompass all of the 94 causal factors. In the first step in this process, pilots from EGL were recruited to participate in a web-based card sorting task. A card sorting program called WebSort™, developed by Dr. Larry Wood at BYU, was used to assess how pilots cognitively organize these human factors issues. The program required each pilot to independently sort the 94 factors or “cards” into categories by placing them in folders and then naming each folder. Each pilot could create as many – or as few – folders as they desired. The pilots could also create “nested” folders (i.e., a single folder could include multiple sub-folders). Because this was a data summarization task, there was only one stipulation: the number of folders must be smaller than the number of factors. Figure 1 below illustrates the card sorting task.

FIGURE 1. WEBSORT™ EXERCISE SCREEN CAPTURE



Pilots in training at EGL were recruited to participate in the WebSort™ exercise. Twenty-seven pilots of varying degrees of experience completed the exercise. The next step was to use this card sort data to develop categories. Card sorting data is typically analyzed via cluster analysis (Ewing, et al., 2001). There are a number of different software programs available for these types of analyses. However, they generally require a distance matrix as input. The following section describes how the card sort data from the 27 pilots were transformed into a single distance matrix.

For each pilot, a 94 x 94 matrix was created. If two items were sorted in the same <u>folder</u>, they received a similarity score of 1. If two items were sorted in the same <u>sub-folder</u>, they received a similarity score of 2. If two items were not in the same folder, they received a similarity score of 0. Again, this analysis yielded a unique matrix for each pilot. Table 1 shows a 5-item example from one pilot. The top half of the matrix is blank because it is redundant with the lower half. Notice that the principal diagonal (the diagonal from the upper-left to the lower right) will always have values of 1 or 2. This is because an item is always going to be paired with itself. The only difference is if it is a regular folder (a value of 1) or a sub-folder (a value of 2).

TABLE 1. EXAMPLE SIMILARITY MATRIX

Item	Item #1	Item #2	Item #3	Item #4	Item #5
Item #1	1				
Item #2	1	2			
Item #3	0	0	2		
Item #4	2	1	1	1	
Item #5	0	0	1	2	1

A composite matrix was then created that summed the corresponding values for each cell (across all the pilots). In this case, A = the sum of the similarity scores for item #1 with itself (across all the pilots). B = the sum of the similarity scores of item #1 with item #2 (again, across all the pilots), and so on for the remaining items. See a sample of Items 1-5 in Table 2 below.

TABLE 2. COMPOSITE MATRIX FOR ITEMS 1 – 5 ACROSS ALL PILOTS

Item	Item #1	Item #2	Item #3	Item #4	Item #5
Item #1	A				
Item #2	B	F			
Item #3	C	G	J		
Item #4	D	H	K	M	
Item #5	E	I	L	N	O

Next, the largest value in this summary matrix was identified, and each cell in the matrix was divided by the value of that cell. For example, if H was the cell with the highest value (item #2 and item #4), then each cell would be divided by the value of H. This created a "distance score" for each cell in the matrix that ranged 0 and 1 (H divided by itself = 1), with larger values representing greater similarity (see Table 3 below for an example).

TABLE 3. MODIFIED COMPOSITE MATRIX FOR ITEMS 1 – 5 ACROSS ALL PILOTS

Item	Item #1	Item #2	Item #3	Item #4	Item #5
Item #1	A/H				
Item #2	B/H	F/H			
Item #3	C/H	G/H	J/H		
Item #4	D/H	H/H	K/H	M/H	
Item #5	E/H	I/H	L/H	N/H	O/H

Since cluster analysis, the statistical technique being used, assumes that smaller values indicate greater similarity (i.e., less distance in multidimensional space), the value in each cell in the matrix was subtracted from 1. This "recoded" the distance matrix so that smaller values represent greater similarity while leaving all values on the same 0 to 1 scale.

The result was a 94 x 94 distance matrix that summarized the information from the 27 pilots who completed the WebSort™ task. This matrix was analyzed using a technique known as hierarchical agglomerative cluster analysis. This method was chosen because its primary function is to sort cards into mutually exclusive groups, which unlike factor analysis, may group cards into more than one group. Simply speaking, cluster analysis looks for the two most closely related items (i.e., based on their distance scores) and groups them together. The software then identifies the next closest item and includes that item within the group. This process continues iteratively until all of the items are included in a cluster. Cluster analysis results are often displayed graphically in a "tree diagram" or "dendrogram."

As previously stated, there are a number of different software programs available for conducting cluster analyses, and they offer a number of different clustering algorithms that determine similarity mathematically. Example algorithms include average linkage method, centroid clustering, complete linkage method, density method, flexible method, McQuitty's method, and median method. Each algorithm produces slightly different results. However, because the cluster analysis results will be used to design a taxonomy that has a menu-type structure, there were certain guidelines that we wished to follow. The criteria for evaluating the categories, and hence the "best" algorithm, are discussed below.

Success Criteria

One criterion for evaluation of classification systems identified in the literature review is utility (Fleishman & Mumford, 1991; see Beaubien & Baker, 2002a, 2002b); using the taxonomy should be easy and efficient. In addition, previous research suggests that most people can only hold seven plus or minus two items in working memory (Miller, 1956). Consequently, in order ensure that the taxonomy was easily manageable, one rule was to ensure that no more than seven to nine high-level categories were created.

Similarly, each category should contain roughly equal numbers of factors or "cards." For example, the menu structure would not be particularly useful if one cluster contained 80% of the factors, and the remaining six clusters shared the remaining 20%. This would overload the user.

Finally, each cluster must make intuitive sense. Quite simply, if the classification system doesn't make sense to the pilots, they won't use it. "Intuitiveness" of the clusters was assessed by looking for common themes in the factors assigned to each cluster.

As previously stated, cluster analysis can be conducted using a number of different clustering algorithms. Each of these algorithms defines mathematical "similarity" in somewhat different ways. Consequently, different clustering algorithms are likely to produce different solutions. As these rules are somewhat arbitrary, there are no formal established rules for which algorithm should be used in which circumstances (Aldenderfer & Blashfield, 1984). Consequently, we began by testing several clustering techniques using the SAS® system, including the average, centroid, complete, density, flexible, McQuitty, and median methods. In order to meet the first criterion for success, we provided instruction in the programming of each algorithm to result in a seven-cluster solution. When instructions were included in the programming code to produce a seven cluster solution, the flexible algorithm produces the most "desirable" result, statistically speaking, in that it produced seven clusters that generally contained 12-25 items per cluster. This result satisfied our second criteria for success. However, upon closer inspection and review, the seven major clusters were found to be virtually un-interpretable. Understandable patterns in the clusters could not be identified, thus violating our third criteria for success. In sum, requiring a seven factor structure using the SAS® system was unproductive.

Although disappointing, this result was not altogether surprising; when instructed to do so, cluster analysis, like many statistical techniques, will arrive at the requested seven cluster solution regardless of whether the data support an intuitive seven cluster grouping (Punj &

Stewart, 1983). To determine how to proceed, a review of the published literature on creating lists of menu items was conducted. This review suggested that the average linkage algorithm is often used for clustering "cards" to develop user interfaces (c.f. Ewing et al., 2002; Toms, Cummings-Hill, Curry, & Cone, 2001). This method is attractive because it associates a factor with a cluster if it is similar on average to the other factors in the cluster. This represents a compromise between the single linkage algorithm, which associates a factor with a cluster if it is similar to only one factor in the cluster, and the complete linkage algorithm, which only associates a factor with a cluster if it is similar to *all* the factors in the cluster (Ewing et al., 2002). Finally, it was discovered that IBM®'s EZCalc software program (Beta Version 1.3) allows one to manipulate the distance score thresholds. This modification allows for the production of clusters with roughly equal numbers of items per cluster, which would assist in meeting our second criteria for success. We proceeded with this average linkage analysis using the EZCalc software.

Note that the data were analyzed three ways in EZCalc:

- "Low-expertise" group;
- "High-expertise" group; and
- all participants together,

where pilots in the high expertise group have an average of approximately 11,000 total flight hours and pilots in the low expertise group average 5,000 total flight hours. Results suggest that all three groups produce similar clusters. Thus it was determined that the combined group of both novices and experts should be used in the analyses to develop the cluster solution. As previously stated, EZCalc allows the user to modify the similarity thresholds, which allows the user to develop clusters with similar numbers of factors. By modifying the similarity threshold slightly from .30 to .37, we were able to produce the nine-cluster solution displayed below in Table 4.

This result was satisfactory in that the number of clusters was reasonable, there were similar numbers of factors per cluster, and the clusters of factors seemed to "go together," thus meeting all three criteria for success.

TABLE 4. AVERAGE LINKAGE ALGORITHM RESULTS

	Cluster								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Number of Factors	5	9	2	19	13	24	5	5	12
Issue Captured	Weight & balance issues	Procedural issues & deviations	Errors made by other people	Pilot error	Organizational factors	Crew resource management & physiological factors	Equipment limitations & failures	Weather	Unexpected Events

Next, we sought to test the validity of the nine-category solution. Three senior pilots from EGL were asked to review the cluster analysis results. Their comments were extremely similar. Most significantly, they suggested reducing the number of clusters from nine to seven. They identified the following clusters:

- Policies or procedures
- Human error
- Human factors
- Organizational factors
- Hardware
- Weather or Environment
- Airspace or ATC

The three EGL pilots then identified what they considered to be the correct assignment of each factor to a cluster. This assignment scheme was deemed the "gold standard."

In sum, the result of Phase I was a comprehensive list of 94 causal factors, a high-level categorizing scheme that contained seven clusters, or categories, and a SME approved assignment of factors to these categories. The goal of Phase II was to determine if pilots could reliably re-classify the 94 factors into the seven categories according to this "gold standard."

Phase II

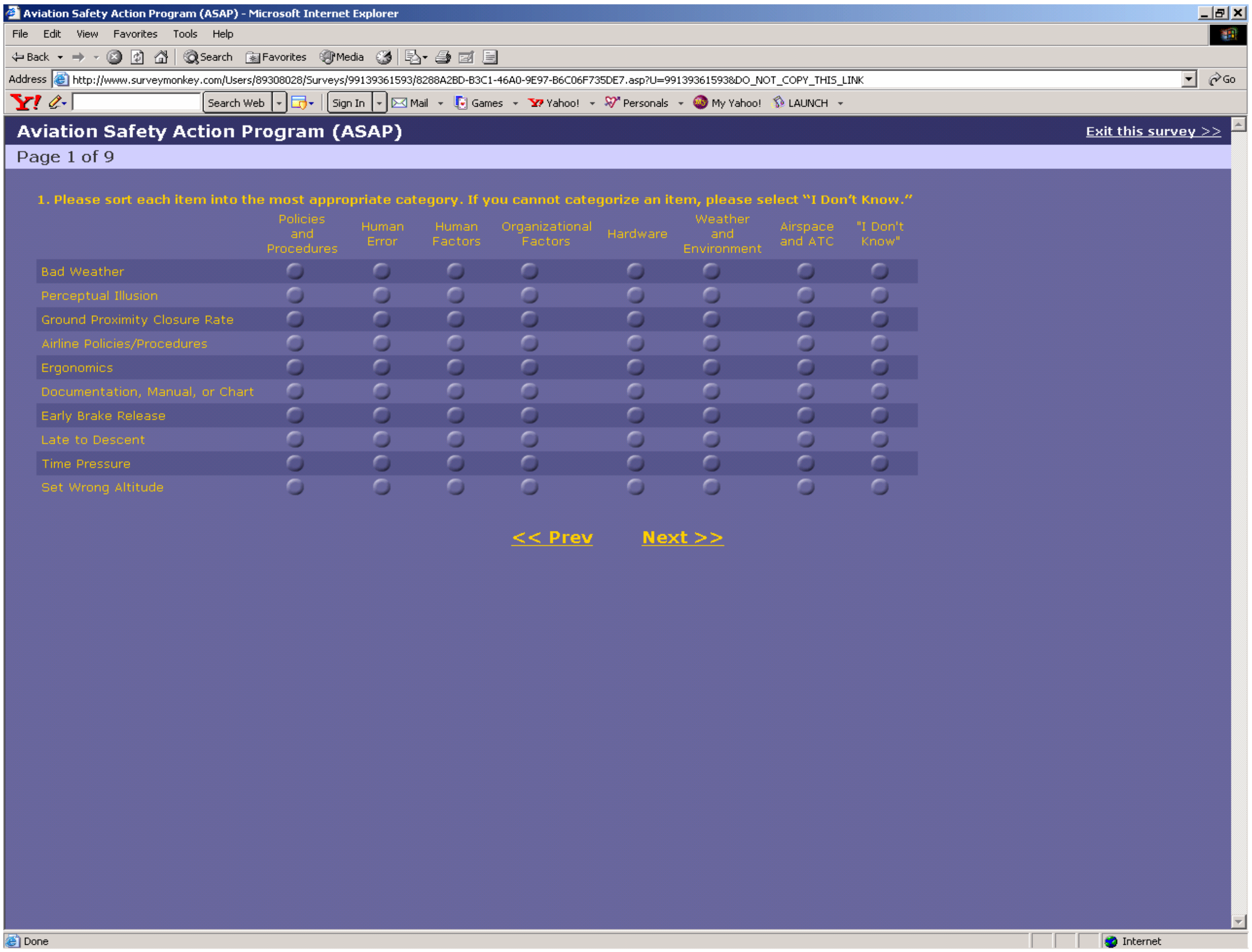
Reclassification Task

The objective of Phase II was to discover if pilots could reliably sort the 94 causal factors into the same categories as the existing "gold standard." Demonstrating reliability at the factor level would further demonstrate internal validity of the taxonomy. To investigate this, an online survey was created and administered using the SurveyMonkey© service (www.surveymonkey.com). The online survey included an introduction page that provided the name and definition of the seven categories, and nine additional pages each containing subsets of the causal factors. Participants were instructed to sort each factor into one of the seven categories. The factors were randomized both across and within pages to prevent order effects. See Figure 2 below for a screen capture of the first page of the classification task in the Survey. See Appendix A for the entire survey.

EGL pilots were invited to participate in the web-based survey. By early 2004, over 100 EGL pilots had responded. The pilots' responses were compared to the gold standard on an item-by-item basis. Success was determined by the percent of agreement between the pilots' responses and the existing gold standard (see Table 5). Specifically, if 75% or more of the pilots agreed with the gold standard, the factor assignment was considered acceptable, and the gold standard validated for that factor. Twenty-two factors were deemed acceptable by this method as they were assigned to the gold standard category by 75% or more of the pilots participating. These factors appeared to be successful because they were specific and were in some way consistent with the category name (e.g., the factor name shared a word in common with the category name). For example, *Equipment Malfunction* was consistently assigned to the *Hardware* category. Similarly, *Organizational Culture* was consistently assigned to the *Organizational Factors* category.

The remaining factors were similarly grouped based upon agreement with the gold standard. If 51-74% of the pilots agreed with the gold standard assignment for a factor, the factor assignment was also deemed to be acceptable. Thirty-four factors met this standard. Factors in this category were sorted into the "correct" category by the majority of individuals but a significant number (up to 50%) assigned it to some other category. It appears that at least in some cases, this may have been due to a lack of clarity in the factor, and a likely subsequent lack of understanding of how the factor might contribute to an incident. For example, the *Night Visual Approaches* factor was assigned by 52% of SMEs to the *Weather and Environment* Category, which is the gold standard. However, 48% assigned it to a different category as follows: *Human Factors* (28%), *Human Error* (5%), *Policies and Procedures* (7%), *Airspace and ATC* (3%) and *I Don't Know* (4%). Although definitions were provided for the categories, factor level definitions were not provided in the survey. SMEs might simply have not understood the factor.

FIGURE 2. PHASE II WEB-BASED SURVEY SCREEN CAPTURE



Note that although these 34 factors were deemed acceptable, given the relatively low level of agreement with the gold standard, it was anticipated that these items might at some later time need to be reworded for clarity, or perhaps moved to a different category, likely during usability testing in Phase III.

Finally, if less than 50% of the pilots agreed with the gold standard assignment for a factor, the factor assignment was deemed unacceptable. Thirty-five factors were deemed unacceptable by this method as they were assigned to the gold standard category by 50% or less of the pilots participating. Many of these factors deemed as unacceptable appear to be the result of confusion between process and outcome. That is, in many cases SMEs assigned an outcome type measure (e.g., *Late to Descent, Failure to Use Autopilot*) to the *Human Factors* category instead of the *Human Error* category. Based on these difficulties, these factors need to be reanalyzed.

TABLE 5. PHASE II RESULTS FROM WEB-BASED SURVEY

Assignment of Factor to Category	# of Factors
Acceptable: 75% or higher agreement	22
Acceptable: 51-74% or higher agreement	34
Unacceptable: 50% or less agreement	35

*Note that the number of factors sums to 91. This is due to an oversight in Phase I, where several factors were inadvertently repeated. The reduction in the number of factors here represents the deletion of the repeated items, not a substantive change in the taxonomy.

In sum, Phase I resulted in a list of factors sorted into seven categories. However, in the first step of Phase II, SMEs were unable to reliably sort approximately 1/3 of the factors into one of the seven categories. The next step was for EGL SMEs to review the unacceptable factors and make a determination regarding their proper category assignment.

Review and Resolution of Unacceptable Factors

The goal of this step was to resolve the factors that were not consistently sorted into the appropriate categories by pilots in the reclassification task. In June 2004, three senior level EGL SMEs from the safety office reviewed the discrepancies between the previously established gold standard and the assignment made by the majority of the pilots responding to the web-based survey. The SMEs made independent decisions regarding assignment of the factors. The three SMEs then convened and discussed to consensus. The result was a decision regarding the most appropriate assignment of the factor to a category, thus resolving the assignment of these factors.

The finalization of the factor assignments to a category concluded Phase II. The result is a taxonomy that organizes contributors to error in aviation performance, and whose development was guided by sound evaluation criteria including internal validity and usability.

Phase III

Taxonomy Reliability and Usability Assessment

Having developed a high-level seven-category structure for the taxonomy, and having determined how each of the 91 factors should be assigned to those seven categories, the goal of Phase III was to assess the taxonomy's overall reliability and usability. Phase III tested these issues and refined the taxonomy as part of an iterative process. The taxonomy was tested by SMEs, and as a result of the data gathered, was modified and tested again. Four rounds of taxonomy review and modification were conducted in Phase III. Table 6 provides an overview of the general sequence of events. Specifics of this process (Steps 1-4) follow.

TABLE 6. PHASE III OVERVIEW

Step	Date	Activity	Attendees
1	6/24/2004	ASAP Meeting, EGL, DFW Airport	N= 7 SMEs*
2	7/15-16/2004	Telephone Conference	EGL SMEs
3	8/5/2004	ASAP Meeting, EGL, DFW Airport	N = 6 SMEs*
4	8/17/2004	Telephone Conference	EGL SMEs

*ASAP meetings are generally attended by representatives from the airline's safety office, the FAA, and the Air Line Pilots Association, Int'l. Included in this count is an analyst from EGL who participated in both ASAP meetings. However, because the analyst is not a pilot, the analyst's taxonomy assignments were not included in the analyses.

Step 1.

On June 24, 2004 an AIR representative attended an EGL ASAP meeting at the EGL safety office in Texas. The purpose of EGL ASAP meetings is to review de-identified ASAP reports and to make various group-level determinations regarding the incidents including determining the level of risk and what might be done to prevent such incidents from reoccurring. In addition to this general purpose, ASAP committee members attending this meeting were also asked to test the use of the taxonomy. Members were provided with a brief introduction to the taxonomy research including information regarding the sponsor and how the taxonomy could ultimately be used in ASAP reporting systems. Members were then provided a copy of the draft taxonomy, which included the seven high-level category labels and their definitions, and a table showing the assignment of each of the 91 factors to its corresponding category. They were also provided a rating form on which they were instructed to record their taxonomy assignments.

Members first participated in a practice run during which they independently read several de-identified reports and then used the taxonomy to identify or "code" the causal contributors. That is, after reading the report, SMEs selected a causal category from the seven high-level categories in the taxonomy that they believed was the primary reason for the incident (the primary category). Then, they selected a causal factor from among the factors associated with that category (primary factor). Next, they repeated the process for what they believed was secondary contributor (secondary category and secondary factor). Members were instructed to make secondary assignments only if they believed that a secondary cause was warranted based n the text narrative in the report.

After each taxonomy assignment was made, the members discussed their decisions as a group until they reached consensus before moving on to the next report. Once the practice run was complete, members then independently coded 44[2] de-identified ASAP reports using the taxonomy. The results were examined for inter-rater agreement (reliability). Results generally suggested that the taxonomy may need additional modification, that it is easier to obtain member agreement at the category level than at the factor level, and that there are patterns to the discrepancies in the assignment. Specifics regarding these results will now be discussed.

To begin our investigation regarding agreement among the assignments that the members made using the taxonomy, we first computed percent agreement at the category level for each report. For example, if four out of the six members selected the same causal category for a report, then the percent agreement for that report was calculated at 67%. We then averaged the percent agreement across reports. The results from these analyses suggest that the taxonomy is quite reliable but not perfectly so. While 100% member agreement was reached in the primary category for 12 of the 44 ASAP reports analyzed from the June meeting (27%), the average percent agreement among the six members was 70% for the primary category. Percent agreement data are presented in the second column of Table 7.

The data regarding secondary assignments is somewhat unique in that members were instructed to assign a secondary category only if they believed a secondary cause existed based on the text narrative provided in the report. As a result, much of the secondary data is "missing." Consequently, when calculating percent agreement for secondary categories, we only included reports for which at least 60% of the members provided an assignment (i.e., believed there was a secondary cause). Analyses conducted on the reports that met this criterion identified 100% agreement for only three of 15 ASAP reports analyzed (20%). However, the average percent agreement among these reports was 69%. Despite the significant reduction in the number of reports available to analyze at the secondary level, the results are quite similar to the primary category assignments.

Next, we investigated percent agreement at the factor level. Note that members were said to agree at the factor level only if they selected the same high-level category assignment. Somewhat less agreement was found at the factor level, with average agreement among SMEs at

[2] SMEs actually rated 54 reports, but 10 were deleted from analysis because multiple reports were submitted by the same crewmember.

50% for the primary factor. Secondary factor assignments were subject to the same criterion as secondary category assignments; that is, only reports for which at least 60% of the members provided a secondary assignment were included in the analyses. Using this criterion, percent agreement for the secondary factor was 36%.

Note that occasionally two crewmembers submit an ASAP report for the same incident (e.g., Captain and First Officer submit a report on the same incident). During our ASAP meeting, members read and made assignments on *all* the de-identified reports, even if they were from the same incident. No effort was made to "select" one report over another when overlap existed. We chose this process for two reasons. First, this approach mirrors how the taxonomy will be used once it is embedded in an ASAP reporting system; that is, each crewmember who submits an ASAP report will be required to make a causal assignment using the taxonomy. Second, ASAP reports from the same incident are often very similar, but in cases where they were different, it would be impossible to determine which reported "version" of the incident was the most accurate. It should be noted that to the extent that multiple reports submitted on the same incident are very similar, this "overlapping" of reports may somewhat inflate the data reported in Table 7.

TABLE 7. STEP 1 RATER AGREEMENT – JUNE 2004

Type of Assignment	% Agreement (# of reports)	Multi-Rater Agreement Complete Record Analysis kappa, *p*, (# of reports)	Multi-Rater Agreement Blanks as a Category kappa, *p*, (# of reports)
Primary Category	70% (n=44)	.44, p<.0001 (n=36)	.40, p<.0001,(n=44)
Primary Factor	50% (n=44)	‡	‡
Secondary Category	69% (n=15†)	.15, p<.0003 (n=12†)	.21, p<.0001,(n=44)
Secondary Factor	36% (n=15†)	‡	‡

† Six SMEs were asked to determine the primary contributing cause (category and factor) for each report, but were instructed to provide a secondary cause (category and factor) only if they believed one existed based on the text narrative of the report. This resulted in a smaller number of secondary reports. Use caution in interpreting these results.
‡The shrinkage in factor assignments was too severe to yield meaningful data with the MAGREE macro.

Percent agreement is intuitive, easy to calculate, appropriate for nominal level data, and can be computed even when data are "missing." The results of the percent agreement analyses are quite good; EGL expressed pleasant surprise at the level of agreement realized. However, percent agreement does not take into account the fact that some agreement would have occurred by chance alone. One statistic that attempts to measure agreement of nominal level data above and beyond what one could expect to get by chance alone is kappa (Cohen, 1960). In order to gain a more complete picture regarding rater agreement, kappa for multiple raters was computed using the MAGREE macro (SAS®).[3] MAGREE (Version 1.0) computes kappa statistics based on generalized kappa (Gwet, 2002). Interpretation of kappa values is as follows: kappa equals 1.0 when there is complete agreement, zero when the agreement can be accounted for purely by chance, and negative when agreement is less than chance agreement. Results of these analyses are presented in the third and fourth columns of Table 7.

One requirement of the MAGREE macro is that the number of raters must be the same across all reports; in essence, missing data are not allowed. Consequently, kappa was first computed only on complete records; any report containing a missing assignment by any one of the six raters was eliminated from the analysis. These results are presented in column three. The overall kappa for the six raters across complete ASAP reports is .44 ($p < .0001$) for the primary category assignments (n=36) and .15 ($p < .0003$) for the secondary category assignments (n=12). Although moderate in magnitude, these results are both significant at $p < .0005$, suggesting that the agreement among raters is greater than would be expected by chance.

As stated previously, MAGREE requires complete records (the number of members must be the same across all reports), and factor level agreement can only be calculated for members who agreed at the category level. These requirements placed stringent burdens on the dataset; the shrinkage was too severe to compute kappa via MAGREE at the factor level.

Considering only complete records results in many reports not being analyzed. In order to investigate the impact of missing data on the multi-rater agreement, missing assignments were replaced with the letter "H" to represent a separate category and kappa was recomputed on all records. These data are presented in column four of Table 7. The overall kappa for the six raters across all ASAP reports is .40 ($p < .0001$) for the primary category assignments and .21 ($p < .0001$) for the secondary category assignments. These results are similar to the complete records analysis: kappa is significant at $< .0001$, suggesting that the agreement achieved among raters is greater than would be expected by chance alone. As before, shrinkage in the factor-level assignments was too severe to yield meaningful data with the MAGREE macro and consequently, kappa was not computed at the factor level.

Note that this analysis, which treats nonresponse as a category unto itself, is likely more meaningful for the secondary assignments where members were instructed to leave an assignment blank if they felt that the report did not support a secondary cause. In this case, the "agreement" that there was no secondary contributing cause is useful in terms of assessing interrater agreement and hence understanding kappa. The improvement of kappa from .15 in the

[3] The SAS® Institute provides the MAGREE macro for computing multi-rater agreement as a service to its users, but the macro is not currently part of its commercially marketed statistics software packages.

complete records analysis to .21 using suggests that members did agree, to some extent, on whether secondary assignments were appropriate or necessary across reports. This "treat blanks as a category" analysis is likely not as useful for primary categories where blanks have no obvious meaning and are likely to be random.

Taken together, these results suggest that members did agree, and at a level higher than chance alone, regarding the assignment of cause to incidents. The results also suggest that it is easier to obtain agreement with regard to high-level categories than at the more detailed factor level. Agreement at the category level was quite acceptable, but factor level agreement was lower, at least when computed via percent agreement. Specifically, the average agreement among members for category level assignments (70% and 69% for the primary and secondary category respectively) was higher than the average percent agreement at the factor level (50% and 36% for primary and secondary factors respectively). There are several likely reasons for this result. First, SMEs were provided with detailed definitions of the categories, but were not provided definitions of the factors. SMEs may simply have had differing ideas regarding what the factors meant, thus reducing their agreement.

Second, percent agreement was not calculated for the secondary category and factor when fewer than 60% of SMEs believed that the report contained information to substantiate the assignment of a secondary category and factor.

Third, the results suggest that there is a pattern to discrepancies between SME assignments. Specifically, SMEs appear to have difficulty with the *Human Error* and *Human Factor* categories. That is, when evaluating complete records, 20% (nine out of 44 reports analyzed in Step 1), some SMEs assigned *Human Error* as the primary category and *Human Factors* as the secondary category, while the majority of the remaining SMEs assigned them in reverse. Statistical analyses were conducted to investigate this issue further. In addition to computing the overall level of agreement among multiple raters, the MAGREE macro computes level of agreement at the response level (in this case, by category). The kappa values for primary and secondary assignments are presented in Table 8. Note that when considering complete records, primary assignments of the *Human Error* and *Human Factors* are by far the lowest of all the response categories at .17 (p<.0001) and .09 (p=.0223), respectively. This suggests that members have trouble distinguishing between these two categories.

In addition to analyzing data from complete records, kappa statistics were also computed for primary and secondary category assignments using a dataset that contains all records, but where the blanks were treated as a new category. These results are presented for comparison in the third column of Table 8. The results are slightly different. However, aside from the Blank category that has a kappa of .07 (p=.0353), the *Human Error* and *Human Factors* categories still generate the lowest rates of agreement among members at .18 (p<.0001) and .13 (p<.0006) for the primary category.

TABLE 8. STEP 1 KAPPA STATISTICS BY CATEGORY – JUNE 2004

Category	**Complete Record Analysis** --------------------------------------- **kappa, *p*, (# of reports), Primary** **kappa, *p*, (# of reports), Secondary**	**Blanks as a Category** --------------------------------- **kappa, *p*, (# of reports), Primary** **kappa, *p*, (# of reports), Secondary**
Policies or Procedures	.29, p<.0001 (n=36) -.06, p=.7850 (n=12†)	.25, p<.0001 (n=44) .03, p=.2095 (n=44)
Human Error	.17, p<.0001 (n=36) .12, p=.0573 (n=12†)	.18, p<.0001 (n=44) .15, p<.0001 (n=44)
Human Factors	.09, p=.0223 (n=36) .10, p=.0795 (n=12†)	.13, p=.0006 (n=44) .19, p<.0001 (n=44)
Organizational Factors	.43, p<.0001 (n=36) .37, p<.0001 (n=12†)	.38, p<.0001 (n=44) .25, p<.0001 (n=44)
Hardware	.78, p<.0001 (n=36) -.01, p=.5749 (n=12†)	.70, p<.0001 (n=44) -.02, p=.7249 (n=44)
Weather or Environment	.39, p<.0001 (n=36) -.01, p=.5749 (n=12†)	.39, p<.0001 (n=44) -.01, p=.5777 (n=44)
Airspace or ATC	.71, p<.0001 (n=36) .37, p<.0001 (n=12†)	.73, p<.0001 (n=44) .21, p<.0001 (n=44)
Blanks		.07, p=.0353 (n=44) .31, p<.0001 (n=44)

† Six SMEs were asked to determine the primary contributing cause (category and factor) for each report, but were instructed to provide a secondary cause (category and factor) only if they believed one existed based on the text narrative of the report. This resulted in a smaller number of secondary reports. Use caution in interpreting these results.

In addition to computing agreement among members regarding their causal assignments, members were also queried regarding the taxonomy's usability. That is, after classifying each ASAP report using the taxonomy, members were then asked to rate the taxonomy's usefulness and usability on a 1-5 scale where *1 = The taxonomy was extremely easy to use* and *5 = The taxonomy was extremely difficult to use.* Average ease of use reported by members was 2.57.

It is difficult to assess whether these agreement statistics and ease of use ratings are acceptable. There is no clear standard from other incident reporting programs regarding inter-rater agreement or average ease of use; existing taxonomies provide too little information in order to make such a comparison. The only numeric evaluation of reliability found in the literature review was conducted by the developers of HFACS, who reported a kappa value of .71 for two raters (Wiegmann & Shappell, 2001). However, the raters were rating causal factors of aviation accidents that had been thoroughly investigated and closed by the National Transportation Safety Board, not unsubstantiated uninvestigated text narratives from incident reports. In addition, there were only two raters (an aviation psychologist and a commercial pilot).

In sum, there are no standards for existing taxonomies against which to compare these results. While solid agreement was found, the agreement was not perfect and the ease of use ratings were not as high as was hoped. Taken together, these results and the discussion that was generated during this meeting suggest that there is room for improvement. Specifically, SMEs were asked during this meeting to make their ratings independently. However, as previously stated, before the process began, SMEs reviewed several reports and made their assignments as part of a “practice run.” Discussion generated during these practice assignments yielded a great deal of pertinent information. In addition, after the practice runs were complete and as the meeting progressed, SMEs continued to ask questions for clarification, providing additional clues. Specifics of this discussion follow.

Users reported initial satisfaction with the seven categories. However, as the meeting continued, it became clear that some confusion existed regarding the meaning of the categories. For example, SMEs were unclear whether the *Organizational Facto*rs category included only airline-specific factors or FAA factors as well. Discussion was also generated regarding the *Policy and Procedure* category. Users were unclear whether this category included problems with the policies themselves or users not following policy and procedure. Two reasons are likely for this confusion. First, the category definition provided to SMEs for the *Policy and Procedure* category does not make it clear whether it is the policy or procedure itself or the person’s use/nonuse that is of interest. Second, the category contains both factors that are attributable to the person (e.g., *Non-adherence to policy/SOP*) as well as policy specific problems (e.g., *Conflicting Procedures*).

In addition to these concerns with regard to the high-level categories, SMEs also seemed to have difficulty with the more detailed factors. Specifically, there was some overlap of factors across categories, which created confusion. For example, the time pressure concept appears in more than one category (*Human Error, Human Factors*, and *Organizational Factors*), depending on how it is worded. In addition, some of the factors do not seem to be assigned to the correct category. For example, *Early Brake Release* appears under the *Policy and Proce*dure category, but in fact is a failure to follow policy and more appropriately falls under the *Human Error* Category. It is possible that this factor was assigned to this category because, as previously stated, the category definition was unclear; SMEs participating in early card sorting procedures were also likely confused. Finally, the level of specificity in the factors varies somewhat within categories. For example, the *Human Error* category included *Failure to Use Autopilot* as well as *Misinterpretation of Information*.

In addition to the lack of standards against which to compare these results, another important caveat regarding these results should be noted. Because the members who participated in this assignment process were not present when the incident occurred, they had to make their decisions regarding the causal contributors to the incident based solely on the text narrative in the de-identified ASAP reports. These narratives often contain little detail, and the members were forced at least in some cases to guess or otherwise make inferences regarding the specifics of the situation. This third party problem was a known factor; as previously noted, this is in fact how these causal assignments are usually made (i.e., an analyst reviews the report and assigns cause). Consequently, 100% agreement among the meeting attendees was not expected. Fortunately, this exercise did yield specific issues that were discussed and easily resolved with EGL SMEs in Step 2 discussed below. It should be noted that the implementation of the taxonomy will eliminate this third party problem. The new system will require the submitting crewmember to use the taxonomy to assign cause.

Step 2.

Cluster analysis was never designed to impose a single structure on the data. Rather, it was designed to demonstrate how pilots organize the issues as they are written, thus serving as a guide for developing a taxonomy. Consequently, we realized that changes to the taxonomy could and should be made – for example by moving items from one cluster to another – along the way to make it more user friendly. In addition, our goal from the beginning of the project was to include both empirical and theoretical support for the taxonomy. Consequently, a telephone conference was held over a two-day period in July 2004 with senior level EGL SMEs from the safety office. The issues raised during the June 24, 2004 ASAP meeting (detailed above) were discussed until consensus was reached. Changes that were implemented to reduce confusion and increase usability of the taxonomy included modifying the definitions of the several of the causal categories, and modifying the factor assignments as appropriate. These issues and their resolutions are detailed in Table 9 below.

TABLE 9. STEP 2 SUMMARY OF ISSUES RESOLVED

Issue	Resolution
Users suggested adding a *Hear Back/Read Back* factor (Note: this factor was added before the 6/24/04 assignments were made)	*Hear back/Read back* formally added as a causal factor under the *ATC* category.
Users unsure about the human in the *Human Error* category	Modified the taxonomy to include a query to determine position of person submitting the report (e.g., pilot, flight attendant, etc.).
There is overlap in the *Time Pressure* factor in the *Human Error* category and *Production Quotas* in the *Human Factor* category	Changed Time Pressure factor in the Human Error category to Self-Induced Time Pressure Changed Production Quotas in the Human Factor category to On Time Performance Pressures
Users confused about whether *Hardware* category includes aircraft hardware only.	Modified *Hardware* category factors to indicate whether factor was related to aircraft hardware or ground equipment hardware.
There is overlap between the *Attitudes Towards Safety* factor in the *Human Factors* category and the *Airline's Safety Culture* factor in the *Organizational Factors* category.	Modify the *Attitudes Towards Safety* factor in the *Human Factors* category to *Personal Attitudes Towards Safety.* Modify the *Airline's Safety Culture* factor in the *Organizational Factors* category to *Airline's Safety Culture*.
Late ATC Clearance factor and *Late Runway Change* factor under the *ATC* category are very similar.	*Late ATC Clearance* factor and *Late Runway Change* factor were combined to *Late ATC Clearance*.
The *Failure to Program FMC* factor in the *Human Error* category only applies to some users/planes/airlines. In addition, FMC isn't the only equipment that could be used improperly.	Combined *Failure to Program FMC* factor and the *Failure to Use Autopilot* factor in the *Human Error* category to *Improper Use of Autopilot/FMS/FMC/Nav Equipment*
Users are unsure about what the *Hardware* category includes.	Modified the definition of the *Hardware* category to include ground equipment.
Category definitions include examples that have been eliminated from the taxonomy.	Examples provided as part of the category definitions were updated to reflect the most pertinent factors that are still included in the taxonomy.
An issue was raised regarding whether the taxonomy should include a factor entitled *Mechanic Error* and to which category it should be assigned.	Modified the taxonomy to include a query to determine position of person submitting the report (e.g., pilot, flight attendant, etc.).
Cross-use of the *Human Error* and *Human Factors* categories. That is, in many cases, a subgroup of users assigned *Human Error* as the primary category and *Human Factors* as the secondary category, while the remaining users assigned them in reverse.	Began discussion with SMEs regarding the difference between these two categories. *Human Error* primarily reflects the outcome, or the thing that the pilot ultimately did wrong. *Human Factors* should capture process level causal contributors to the final error. Modified the category definitions slightly to better reflect the difference between human error and human factors. Removed *Pilot Error* factor from the *Human Error* category. In the *Human Error* category, combined the *Improper Weights were Loaded* factor and the *Improperly Loaded Cargo* factor to *Weight and Balance Error*

After the July telephone meetings, an updated taxonomy definitions and assignment list was developed that incorporated all the resolutions and was sent to EGL for approval. Once the new version was approved, another test of the taxonomy was scheduled and conducted. The results of this test are described in Step 3 below.

Step 3.

On August 5, 2004, AIR attended a second ASAP meeting at EGL. As before, a practice run was conducted where SMEs read several de-identified incident reports, made their assignments independently using the updated taxonomy, and then discussed the assignments as a group in order to reach to consensus. However, before getting started with the practice run we discussed several issues as a group. Specifically, we discussed the possibility of adding additional factors to the taxonomy that were identified recently as potential additions (e.g., topography, traffic, ceiling, visibility, snow, rain, fog, overcast, VMC/IMC, mixed marginal, crew knowledge, crew memory, calculation or computation, information gathering and distribution, fuel management, preflight inspection, task management or workload distribution, task priority, time management, contingency planning, leadership, social conversation, flexibility, passenger relations, communication, and coordination). After discussion of these factors, the SME team decided only to change the factor *Low Visibility* to *Low Visibility/Ceiling* in the *Weather & Environment* Category and to add a new factor called *Air Traffic Congestion* under the *Airspace and ATC* category.

The meeting then proceeded as in Step 1. The SMEs were easily able to reach agreement on the practice reports. Next, SMEs made independent assignments of 28 de-identified reports and recorded their assignments on a sheet. Finally, SMEs were asked to rate the taxonomy's ease of use.

The same data analysis process was followed for the member's assignments from this meeting as they were from the June meeting. That is, percent agreement was calculated for primary category, primary factor, secondary category, and secondary factor. Then kappa was computed for the primary and secondary category assignments, both overall and at the category level. These analyses will now be discussed.

Although the number of reports reviewed was smaller in this iteration (28 versus 44 reports in Step 1), and the number of SMEs was reduced (five versus six in Step 1) the results of this iteration were improved. There were fewer problems and discrepancies, and fewer questions were raised during the practice assignments. AIR analyzed the assignment data and found improvements in agreement in the assignments of primary cause with percent agreement across raters and reports increasing from 70% to 71% at the category level and from 50% to 52% at the factor level (see Table 10). Percent agreement of the assignment of secondary causal category dropped from 69% in Step 1 to 61% in Step 2, while agreement at the secondary factor level dropped from 36% to 31%. As before, no effort was made to “select” one report over another when more than one crewmember submitted a report for the same incident.

TABLE 10. STEP 3 RATER AGREEMENT –AUGUST 2004

Type of Assignment	**% Agreement (# of reports)**	**Multi-Rater Agreement Complete Record Analysis kappa, *p*, (# of reports)**	**Multi-Rater Agreement Blanks as a Category kappa, *p*, (# of reports)**
Primary Category	71% (n=28)	.42, *p*<.0001 (n=28)	.42, *p*<.0001 (n=28)
Primary Factor	52% (n=28)	‡	‡
Secondary Category	61% (n=17†)	.27, *p*<.0001, (n=7†)	.14, *p*<.0001 (n=28)
Secondary Factor	31% (n=17†)	‡	‡

† Five SMEs were asked to determine the primary contributing cause (category and factor) for each report, but were instructed to provide a secondary cause (category and factor) only if they believed one existed based on the text narrative of the report. This resulted in a smaller number of secondary reports. In addition, when making their secondary category assignments, SMEs did not use all of the categories. Use caution in interpreting these results.
‡The shrinkage in these cells was too to yield meaningful data with the MAGREE macro.

In order to account for the fact that a certain amount of agreement would have occurred by chance alone, the MAGREE macro (SAS®) was used to compute kappa for multiple raters. The overall kappas for the analysis of the complete records are shown in the third column of Table 10. Specifically, kappa for the primary category assignments is .42 ($p < .0001$) and .27 ($p < .0001$) for the secondary category assignments. These results are both highly significant, suggesting that agreement among raters is greater than would be expected by chance alone. Again, the requirements of MAGREE create shrinkage in the primary and secondary factor assignments that was too severe to yield meaningful data with the MAGREE macro.

The next step was to consider the impact of missing data on our analyses. Blanks in the dataset were replaced with the letter “H” to represent a separate category and kappa was recomputed (see column four of Table 10). The overall kappa for the six raters across all ASAP reports is .42 ($p < .0001$) for the primary category assignments and .14 ($p < .0001$) for the secondary category assignments. Again, the results are highly significant, suggesting greater agreement than would be realized by chance alone. As previously stated, due to the nature of secondary assignments and the instruction for SMEs to leave it blank if no secondary cause was apparent, kappa computed via this substitution method is probably more useful in assessing agreement for the secondary category assignment than at the primary category level. Shrinkage in the factor-level assignments was too severe to yield meaningful data with the MAGREE macro.

As before, SMEs were asked to rate the taxonomy's usefulness and usability on a scale of 1-5 where 1 = *The taxonomy was extremely easy to use* and 5 = *The taxonomy was extremely difficult to use*. Average ease of use reported by SMEs improved to 2.20 from 2.57 in Step 1 (note the scale values where smaller ratings represent increased reported ease of use).

Taken together, these results are encouraging. When considering percent agreement, the assignment of primary cause and secondary cause improved slightly, and reported ease of use improved. Although inter-rater agreement in the assignment of the secondary causal category decreased, this is not considered a significant problem due to the nature of secondary assignments. First, situations are sometimes clearly caused by only one factor. In these cases, selecting a high-level category and identifying the more specific factor is relatively easy and results in a complete evaluation of the situation; no other information is required or pertinent. This means that when comparing secondary assignments among raters, a substantially smaller number of reports is considered. This likely makes the ratings less stable. Furthermore, as previously stated, the SMEs making these assignments were not present when the event occurred; they necessarily have to rely solely on the text narrative submitted, which may contain little detail, particularly regarding less important or obvious causal contributors.

The previous usability test conducted in June 2004 suggested that members were having difficulty with the *Human Error* and the *Human Factors* categories. In order to investigate this issue for the current study, kappa was computed at the category level. These results are presented in Table 11. The results suggest that the confusion between these categories is reduced. The kappa values for *Human Error* and *Human Factors* categories improved to .22 and .24 (both at $p<.0001$) for the primary and secondary categories, up from .13, $p=.0005$ and .06, $p=.0644$ respectively in the first trial. However, despite this improvement, the kappa values are not large. The improvement could be due to practice effects. In addition, for a significant number of reports (eight out of 28, or 29%), members again assigned *Human Error* as the primary cause and *Human Factors* as the secondary cause, while the majority of the remaining members assigned the categories in reverse. This response pattern observed in both Step 1 and Step 3, combined with the general notion of what constitutes error suggests that issue warrants further consideration.

In addition to analyzing data from complete records, kappa statistics were also computed for primary and secondary category assignments using a dataset that contains all records, but where the blanks were treated as a new category. These results are presented in the third column of Table 11. Note that the kappa values presented for the primary categories are the same as the analysis on complete records. This is because there were no blanks in the primary category assignments in this trial. This makes it impossible to determine the impact of missing data at the primary category level. However, there were blanks in the secondary assignments. As in the first trial, the kappa values for all the secondary categories when blanks are treated as a category are all quite low.

TABLE 11. STEP 3 KAPPA STATISTICS BY CATEGORY – AUGUST 2004

Category	Complete Record Analysis -- kappa, *p*, (# of reports), Primary kappa, *p*, (# of reports), Secondary	Blanks as a Category --------------------------------- kappa, *p*, (# of reports), Primary kappa, *p*, (# of reports), Secondary
Policies or Procedures	.65, *p*<.0001, (n=28) -.06, *p*=.6939, (n=7†)	.65, *p*<.0001, (n=28) -.04, *p*=.7323, (n=28)
Human Error	.22, *p*<.0001, (n=28) .07, *p*=.2885, (n=7†)	.22, *p*<.0001, (n=28) .07, *p*=.1054, (n=28)
Human Factors	.24, *p*<.0001, (n=28) .59, *p*<.0001, (n=7†)	.24, *p*<.0001, (n=28) .26, *p*<.0001, (n=28)
Organizational Factors	.17, *p*=.0022, (n=28) .07, *p*=.2885, (n=7†)	.17, *p*=.0022, (n=28) .13, *p*=.0152, (n=28)
Hardware	.54, *p*<.0001, (n=28) NA‡	.54, *p*<.0001, (n=28) -.01, *p*=.5479, (n=28)
Weather or Environment	.48, *p*<.0001, (n=28) NA‡	.48, *p*<.0001, (n=28) -.01, *p*=.5958, (n=28)
Airspace or ATC	.72, *p*<.0001, (n=28) .09, *p*=.2294, (n=7†)	.72, *p*<.0001, (n=28) .27, *p*<.0001, (n=28)
Blanks		NA♦ .03, *p*=.3210, (n=28)

† Five SMEs were asked to determine the primary contributing cause (category and factor) for each report, but were instructed to provide a secondary cause only if they believed one existed based on the text narrative of the report. This resulted in a smaller number of secondary reports. In addition, when making their secondary category assignments, SMEs did not use all of the categories. Use caution in interpreting these results.
‡Members did not utilize this category when making causal assignments.
♦Members did not leave any primary category assignments blank.

The issues identified in the ASAP meeting during Step 3 were summarized and brought up for discussion in a telephone conference with EGL SMEs. This discussion and its conclusions are outlined in Step 4 below.

Step 4.

On August 17, 2004, a telephone meeting was held with senior level EGL SMEs from the safety office to discuss the results of the August 5, 2004 ASAP meeting. As before, the remaining issues were discussed until consensus was reached. The issues were resolved by modifying, deleting, or restructuring. Changes that were implemented to reduce confusion and increase usability of the taxonomy included modifying the definitions of the several of the causal categories, and modifying the factor assignments as appropriate. These issues and their resolutions are presented in detail in Table 12 below.

TABLE 12. STEP 4 SUMMARY OF ISSUES RESOLVED

Issue	Resolution
The *Ergonomics* factor in the *Hardware* category was thought to be too generic.	The factor name was modified slightly to *Ergonomics/Poor Product Design, Placement or User Interface*.
The *Human Factors* category is markedly longer than the others.	Redundant factors were deleted (e.g., *Poor Judgment* and *Delayed Reaction Time*) Ambiguous factors were deleted (e.g., *Unresolved Ambiguity*)
Security was highlighted as a possible issue in one report, although the taxonomy does not currently contain a *Security*-related factor.	SMEs indicated that security issues are generally sent to another department to be managed and tracked, and are not officially part of the ASAP program. Consequently, it was recommended that the taxonomy not include a security factor.
Users occasionally switch the *Human Factors* category and the *Airspace and ATC* category (i.e., some users assign primary causation to *Human Factors* and secondary to *Airspace and ATC*, while the majority of remaining users made the assignments in reverse.	It was determined that this could best be resolved via training and making sure the entry form is clear.
The definitions for both the *Human Factors* category and the *Hardware* category suggest that product design issues should be included.	Given that EGL pilots assigned the *Ergonomics/Poor Product Design, Placement or User Interface* factor to the *Hardware* category, this assignment will be retained. However, we modified the beginning of the definition of the *Human Factors* category from "Human Factors influence our interactions with one another and with technology…" to "Human Factors are physical, psychological, and social factors that influence human performance…"
Users continue to mix assignments to the *Human Error* and *Human Factor* categories.	SMEs suggested that users are looking at the error as the cause, instead of looking for the cause of the error. SMEs first indicated that they did not see this as a problem. However, upon further discussion, it was decided that while the taxonomy would include a Human Error category, that when submitting ASAP incidents, crewmembers should be required to identify the error (i.e., the outcome or result) first, and to identify the cause separately.
The need for definitions of the causal factors was discussed. AIR and EGL SMEs agreed that a list of definitions was important.	AIR agreed to develop a draft list of definitions, and EGL agreed to review and make suggestions.

As a result of this final SME review and discussion, the causal category definitions and the assignment of factors to a category were finalized. AIR developed a list of factor definitions, which was submitted to EGL for review and comment. Several rounds of revision were required for AIR and EGL to finalize the factor level definitions.

This concluded the development of Version 1.0 of the taxonomy, which consists of seven high-level categories and 70 associated factors. Upon completion, the taxonomy was officially named Aviation Causal Contributors for Event Report Systems, or ACCERS (see Appendix B).

Summary

Phases I, II, and III were designed to scientifically develop and test the ACCERS taxonomy for making causal attributions in ASAP incident reporting systems. Reliability and validity were built into the taxonomy at each step of development by incorporating the best statistical techniques as well as the opinions of qualified SMEs. The process and results suggest that the taxonomy is a sound and useful tool. It is anticipated that implementation of the taxonomy into an airline's ASAP system will allow for more accurate and timely data collection and analysis.

This vigorous research process resulted in a great deal of specific information, regarding the strengths and weaknesses of ACCERS, both of which should be taken into account during implementation. These issues will be highlighted in the next section.

DISCUSSION

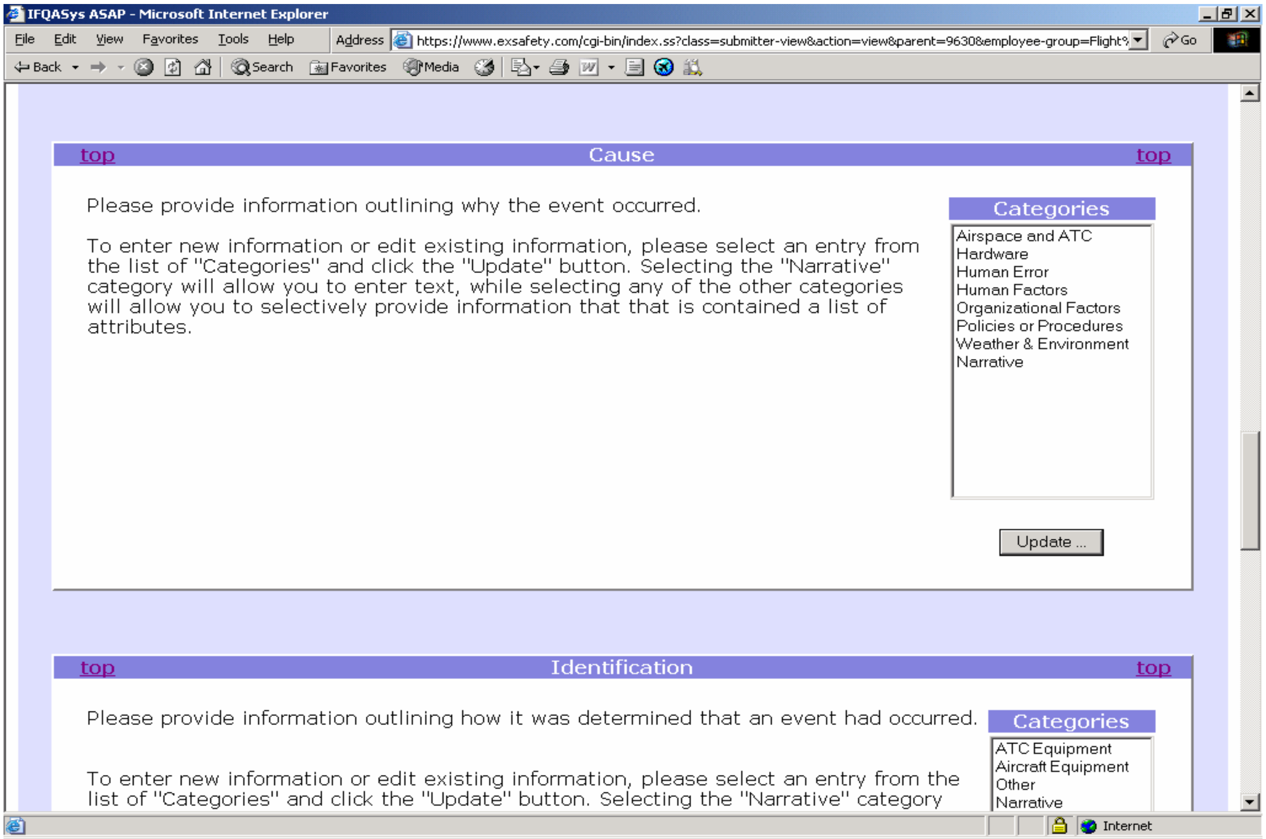
The result of this research project is ACCERS: a taxonomy appropriate for use in classifying causal contributors to human error in ASAP systems. The taxonomy was developed and evaluated based on both theoretical and practical applied considerations, and satisfies many evaluation criteria for classification systems. The next step is to put ACCERS to use solving real problems to enhance safety in real life applications. However, the development and evaluation process yielded a great deal of information regarding taxonomies, their use, and the people who use them. This section outlines some current applications of ACCERS as well as issues to consider for implementation.

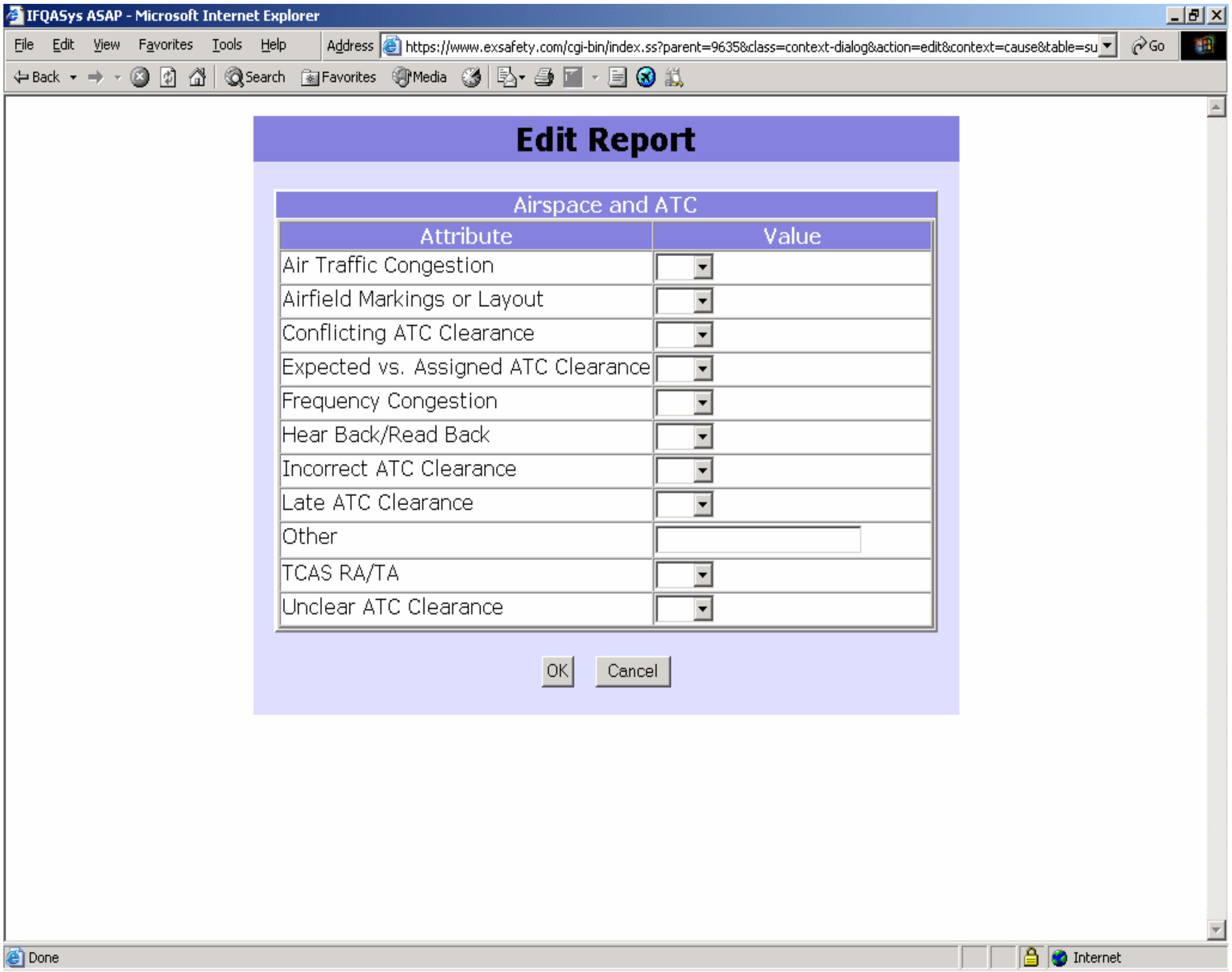
As previously stated, it was our goal for the taxonomy to be an immediately usable product. We are pleased to report that the taxonomy has already been implemented into a web-based ASAP reporting system called the Integrated Flight Quality Assurance System (IFQASys) being built by Universal Technical Resource Services, Inc. (UTRS). Details of this process follow.

The transition of the two systems took place in a series of events. Fist, a teleconference was conducted with AIR, EGL, and UTRS representatives in August, 2004. This meeting revealed a significant amount of overlap between the work being conducted by UTRS and the AIR-EGL team. In October 2004, representatives from EGL, UTRS, the FAA, and AIR met at AIR's in offices Washington, DC to discuss how best to consolidate ACCERS into IFQASys. It was ultimately determined that because UTRS had already developed a comprehensive taxonomy, that it would be best to embed ACCERS into their existing system. As their only direct need was for information regarding human factors as a specific category, the factors in ACCERS's *Human Factors* category were incorporated in their entirety into the design of IFQASys. It was also decided that, because ACCERS was built empirically using SME input, that IFQASys should reflect ACCERS as accurately as possible. Consequently, representatives from AIR and UTRS worked together to verify that the ACCERS factors appeared in IFQASys's design. This process and the few minor modifications needed were completed in November 2004 via telephone and electronic communication between AIR and UTRS.

In December 2004, the IFQASys project and the system itself underwent significant changes. Changes to design of the web-based interface involved moving to a more component-based system. That is, IFQASys now consists of many components including Overview, Contact Information, Employees, Description, Cause, Identification, and Response, which are captured in separate sections of the overall system. ACCERS is no longer embedded in a larger taxonomy but rather exists in full in the "Cause" section of IFQASys. Figure 3 below show screen captures from IFQASys including the main "Cause" screen capture that shows ACCERS's seven high-level categories, and a second screen capture that shows the specific factors associated with one of the categories – the *Airspace and ATC* category. IFQASys opened to contributors in late May, 2005; data are currently being collected.

FIGURE 3. IFQASYS "CAUSE" MODULE SCREEN CAPTURES





The implementation of ACCERS into IFQASys represents a strong start. However, as previously stated, the research involved in developing and evaluating the taxonomy revealed some interesting issues. First, in Phase III, the reliability and ease of use of AIR's taxonomy did appear to improve between Step 1 and Step 3. Although it is not known whether the improvements were due to the modifications made to the taxonomy or to the practice effect (several of the SMEs participating in this Step 3 meeting also participated in the first meeting), training individuals in the appropriate use of the taxonomy is considered to be an integral part of implementation of the tool.

A second consideration stems from the usability tests of the taxonomy, during which SMEs seemed to confuse or crossover their use of the *Human Error* and *Human Factors* categories. It was felt among the AIR team that this assignment problem was a function of two issues. First, the *Human Error* category contains only a small number of errors expressed in high-level terms (e.g., *Misapplication of Flight Controls*, *Weight and Balance Error*). The crossover problem may be a function of the fact that the factors listed do not represent a comprehensive set of human errors likely to be committed by pilots. Second, the AIR team believes the difficulty in discriminating between the *Human Error* and *Human Factors* categories may be because the *Human Error* category generally captures the outcome of performance (the error itself) and the *Human Factors* category generally captures the process of performance (the cause of the error). This is an important distinction in that it is the cause of the error, not simple identification of the error itself that leads to productive prevention strategies.

Based on these considerations, the fact that it is impossible to determine a priori which actions are processes and which are outcomes, (any action can theoretically be a process or an outcome depending on the situation), and based on the literature review that revealed the importance of capturing outcomes as well as contributing factors, it is AIR’s recommendation to remove the *Human Error* category from the taxonomy, and to instead ask the pilot or crewmember to identify the outcome in a separate question from the identification of the causal factors. Using a two-tiered approach, the user will be asked to identify the outcome first, and then will be asked on a separate screen to identify the underlying cause.

Another implementation issue to be considered involves the fact that the taxonomy was developed in part by SMEs from EGL. As such, the selection and definition of various causal concepts make sense in EGL’s culture. However, these assignments may not be appropriate in another organization. For example, the crew resource management (CRM) factors that are currently part of the *Human Factors* category in the taxonomy (e.g., *Incompatible Crew Pairing*, *Conflict Management*) are appropriate for EGL pilots and the EGL culture. However, other airlines may have other CRM priorities.

To address these issues, it is recommended that organizations wishing to use the taxonomy incorporate the established seven-category causal structure, but investigate whether the causal factors are appropriate for use in their organization. For example, it is recommended that organizations continue to assign CRM factors to the *Human Factors* category, but that they identify the appropriate specific CRM factors by assessing their organization’s view on CRM. Reviewing training materials, safety statements, and otherwise identifying the organization’s culture with regard to teamwork would be appropriate methods for identifying the appropriate CRM factors to be included. Empirical investigation could also reveal a great deal regarding whether the taxonomy is appropriate for use or whether it needs to be modified before being implemented.

In addition to evaluating the appropriateness of various factors, some factor assignments may also need to be evaluated before implementation. For example, while AIR and EGL SMEs agree that *Ergonomics* and *Spatial Disorientation* are important factors, there was some disagreement between the EGL and AIR experts regarding to which category these factors should be assigned; EGL SMEs assigned these factors to the *Hardware* category, while AIR SMEs believe that they belong more appropriately in the *Human Factors* category. These differences are likely a function of the difference in our backgrounds and experience; pilots clearly categorize these issues in a way that is different from human factors research analysts. The fortunate discovery of these differences in assignment is a function of our research method: the taxonomy was built by pilots, not researchers. However, differences do exist. It is recommended that in order to avoid confusion stemming from the fact that internal safety analysts who review and analyze the data from ASAP reports might cognitively place these things in a different category than pilots, it might be useful to examine how pilots in the implementation organization categorize them.

Before implementation, it is critical that all users be committed to the definitions of both the categories and the individual factors. Each of these documents should be reviewed and accepted or modified as appropriate.

Finally, a critical improvement in the use of ACCERS is the notion that the crewmember who was present when the incident occurred makes the determination regarding cause. Because the research conducted in this project relied on third party SMEs (experts who are technically savvy but who were not present when the incident occurred), it is highly recommended that organizations who implement the system conduct a pilot test of the taxonomy. During this pilot test, each causal category should include an "Other: [Type the causal contributor here]" option. After the system has gathered data for a period of 3-6 months, safety managers and analysts should review and analyze the data. Specifically, trends in reporting should be examined (e.g., factors that are rarely selected, "Other, Type in" responses that should be added as factors). The information gathered should be used to modify the taxonomy and/or the reporting system.

In sum, while ACCERS was built with empirical input from EGL pilots, the research conducted during the development and evaluation of the taxonomy shed light on a number of pertinent issues regarding its implementation. It seems reasonable that the taxonomy will generalize to other airlines if these recommendations are used in making the transition to the use of ACCERS.

NEXT STEPS

The previous section provides implementation recommendations based on an understanding of the issues raised during the development and evaluation of ACCERS. The recommendations provided are relatively simple and straightforward; it is believed that any organization wishing to implement the system could do so with little effort.

In addition to providing information regarding appropriate implementation strategies, the research conducted to develop and evaluate ACCERS also provides the theoretical underpinnings for future research both with regard ACCERS development and as a way to enhance the voluntary reporting process overall. These research ideas, or next steps, are outlined below.

Taxonomy Development

As previously discussed, Fleishman and Mumford (1991) suggest that utility is an important asset for a classification system. Utility in this context includes the promotion of communication, efficient use of resources, solution to applied problems, and wide acceptance by users. ACCERS now exists in IFQASys and is being considered by other agencies as well. To further enhance the implementation of ACCERS, we believe the development of a code book or training manual is appropriate.

In addition to a code book, it was suggested in the implementation section above that the *Human Error* category should be removed from ACCERS in favor of a two-tiered process/outcome approach. In order for the human error outcome data to be useful, it is anticipated that pilots will need to be provided with a taxonomy of human errors from which to choose. It is possible that the factors identified in this project in the *Human Error* category could be used as fodder for that taxonomy. Although logical in its approach and supported at least in theory, the process/outcome technique has not yet been developed or tested with users. Consequently, development and evaluation of this two-tiered approach and a more complete taxonomy of human error is recommended as a next step in the development of ACCERS.

In addition to the implementation recommendation that organizations review their safety culture with regard to CRM and identify the most appropriate CRM factors, an important next step in the development of taxonomies in general is an investigation of the global nature of CRM constructs. While CRM is generally accepted as an important component of safe cockpit behavior, and a great deal of research has been conducted to understand and define the constructs involved, there is little consensus regarding which CRM constructs are global in nature (i.e., which CRM constructs are important and should be considered in any and all organizations). Synthesis or meta-analysis of the available research could reduce the amount of time and effort required to implement taxonomies, and increase the comparison of data across organizations. Finally, the identification of global CRM constructs may assist other industries (medicine, maritime transportation) in developing their own CRM human factors taxonomies.

Reliability is a key concept in evaluating classification systems. The results of this research suggest that the level of agreement among third party contributors (SMEs who are technically savvy but who were not present at the incident) is greater at the category level than at the factor level. Consequently, it may be useful to further investigate the assignment process by examining how and why individuals make causal attributions and assignments. For example, a group of pilots could be shown a situational vignette of an incident and then asked to make causal inferences using ACCERS. Assignment differences at the category and factor level could be examined and discussed to investigate how individuals make causal attributions and how they make decisions regarding which is the most important causal category or factor. This information could be used to develop training materials for users or to modify ACCERS as appropriate.

Finally, additional reliability evidence could be gained by comparing whether SMEs who are similar in experience and training make the same causal attributions regardless of the length of the text narrative or the level of detail included in the narrative.

Link System Data via the Taxonomy

The development of ACCERS represents a significant step towards the goal of maximizing the utility of available data for the purpose of reducing or eliminating barriers to safety in aviation. In addition to improving the ASAP system, ACCERS, if applied in other areas of aviation performance, could theoretically create a great deal of synergy in terms of enhancing the safety system as a whole. Consequently, an additional goal is to use the ACCERS or some modified version thereof to categorize incidents from other job categories such as dispatch or flight attendants.

It may also be possible to use ACCERS to code de-identified AQP performance ratings and FOQA output. The result would be three separate databases that use a common taxonomy. Although individual records will be de-identified, carriers will be able to identify safety-related problems by triangulation. For example, if a carrier's ASAP reports indicate that non-precision approaches are a problem, their FOQA and AQP data can be analyzed to verify the problem's existence.

SUMMARY

This research and development effort resulted in a simple and easy to use taxonomy that can be used in voluntary reporting systems as a way to categorize information and identify trends. However, the development of the taxonomy is just the first step in enhancing safety. The taxonomy must be used in reporting system data and these trends must be investigated and remedied to the extent possible by safety managers and other organization officials. To this end, this report suggests ways to implement ACCERS, and ways to further develop it and the ASAP system overall.

REFERENCES

Aldenderfer, M. S. & Blashfield, R. K. (1984). *Cluster Analysis*. Sage: Beverly Hills, CA.

Australian Transportation Safety Bureau (2000). *ATSB Annual Review 2000*. Canberra, ACT: Commonwealth of Australia.

Beaubien, J. M. & Baker, D. P. (2002a). A review of selected aviation Human Factors taxonomies, accident/incident reporting systems, and data reporting tools. Technical Report submitted to FAA.

Beaubien, J. M., & Baker, D. P. (2002b). A review of selected aviation Human Factors taxonomies, accident/incident reporting systems, and data reporting tools. *International Journal of Applied Aviation Studies, 2*(2), 11-36.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *XX*(3), 37-46.

Ewing, G., Logie, R., Hunter, J., McIntosh, N., Rudkin, S., Freer, Y., Ferguson, L. (2002). A new measure summarising 'information' conveyed in cluster analysis of cart sort data: Application to a neonatal intensive care environment. In P. Lucas, L. Asker, & S. Miksch (Eds.), *Intelligent Data Analysis in Medicine and Pharmacology* Workshop held during the European Conference on Artificial Intelligence, Lyon, France (2002, July).

Fleishman, E. A. & Mumford, M. D. (1991). Evaluating classifications of job behavior: A construct validation of the ability requirement scales. *Personnel Psychology*, *44*, 523-575.

Gwet, K. (2002). Computing inter-rater reliability with the SAS® System. *Statistical Methods for Inter-Rater Reliability Assessment*, 3, 1-16.

Pounds, J., & Isaac, A. (2002). *Development of an FAA-EUROCONTROL Technique for the Analysis of Human Error in ATM*. Washington, DC: Federal Aviation Administration Office of Aerospace Medicine Technical Report No. DOT/FAA/AM-02//12.

Lievens, F. (2000). Development of an empirical scoring scheme for situational inventories. *European Review of Applied Psychology*, 50(1), 117-124.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.

Punj, M. & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, XX*, 134-148.

Reason, J. (1990). *Human error*. New York: Cambridge University Press.

Toms, M., Cummings-Hill, M., Curry, D., & Cone, S. (2001). Using Cluster Analysis for Deriving Menu Structures for Automotive Mobile Multimedia Applications. Reprinted from Human Factors in Automotive Design. SAE 2001 World Congress, Detroit, Michigan, March 5-8, 2001.

Kohn L. T., Corrigan J. M., Donaldson, M. S. *To err is human*. Washington, DC: National Academy Press; 1999.

Wiegmann, D. A. & Shappell, S. A. (2001). Human error perspectives in aviation. *The International Journal of Aviation Psychology, 11*(4), 341-357.

APPENDIX A: WEB SURVEY

Aviation Safety Action Program (ASAP) - Microsoft Internet Explorer

Address http://www.surveymonkey.com/Users/89308028/Surveys/99139361593/EEE8D59F-DDED-40AD-87FD-C1016B84B5F7.asp?U=99139361593&DO_NOT_COPY_THIS_LINK

Aviation Safety Action Program (ASAP) Exit this survey >>

Welcome!

Dear American Eagle Pilot,

As you may know, American Eagle is re-designing our Aviation Safety Action Program (ASAP) reporting system. When complete, you will be able to submit your ASAP reports online.

As part of this project, we are developing a custom menu structure for the ASAP software. Please help us by completing this survey, which will ask you to sort items into categories. The entire survey should take about 20 minutes.

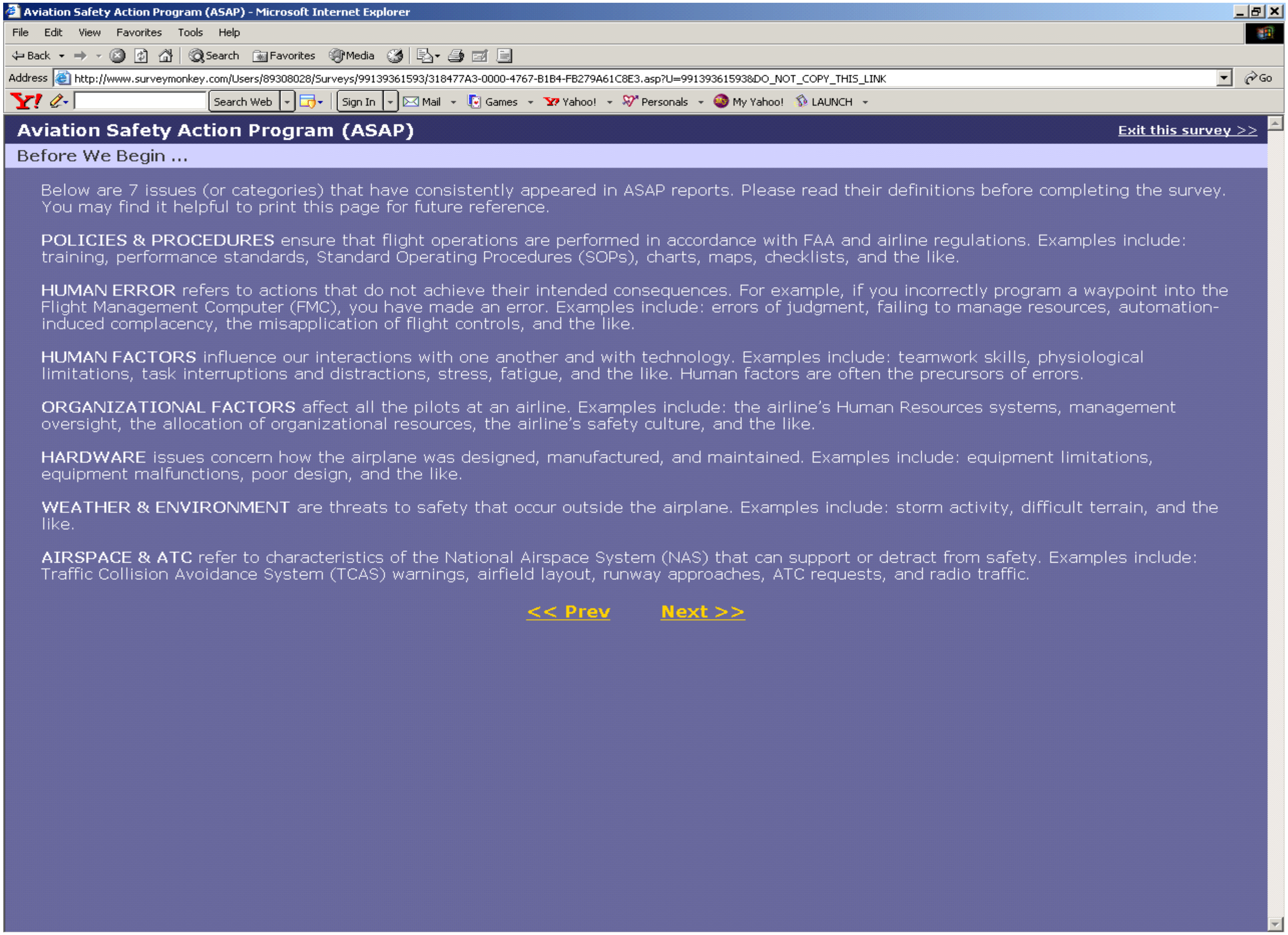
If you would like additional information about this project, please call Capt. Craig Chilcote or Dr. Jeff Beaubien. If you experience computer problems while completing the survey, please call Dr. Jeff Beaubien. During beta testing, we found that the survey is best viewed with a monitor resolution of **1024 x 768** or greater.

Thank you,

Capt. Craig Chilcote
Manager, ASAP Program
American Eagle Airlines
(972) 425-1956

Dr. Jeff Beaubien
Senior Research Scientist
American Institutes for Research
(202) 342-5133

Next >>



Aviation Safety Action Program (ASAP) - Microsoft Internet Explorer

http://www.surveymonkey.com/Users/89308028/Surveys/99139361593/318477A3-0000-4767-B1B4-FB279A61C8E3.asp?U=99139361593&DO_NOT_COPY_THIS_LINK

Aviation Safety Action Program (ASAP)

Exit this survey >>

Before We Begin ...

Below are 7 issues (or categories) that have consistently appeared in ASAP reports. Please read their definitions before completing the survey. You may find it helpful to print this page for future reference.

POLICIES & PROCEDURES ensure that flight operations are performed in accordance with FAA and airline regulations. Examples include: training, performance standards, Standard Operating Procedures (SOPs), charts, maps, checklists, and the like.

HUMAN ERROR refers to actions that do not achieve their intended consequences. For example, if you incorrectly program a waypoint into the Flight Management Computer (FMC), you have made an error. Examples include: errors of judgment, failing to manage resources, automation-induced complacency, the misapplication of flight controls, and the like.

HUMAN FACTORS influence our interactions with one another and with technology. Examples include: teamwork skills, physiological limitations, task interruptions and distractions, stress, fatigue, and the like. Human factors are often the precursors of errors.

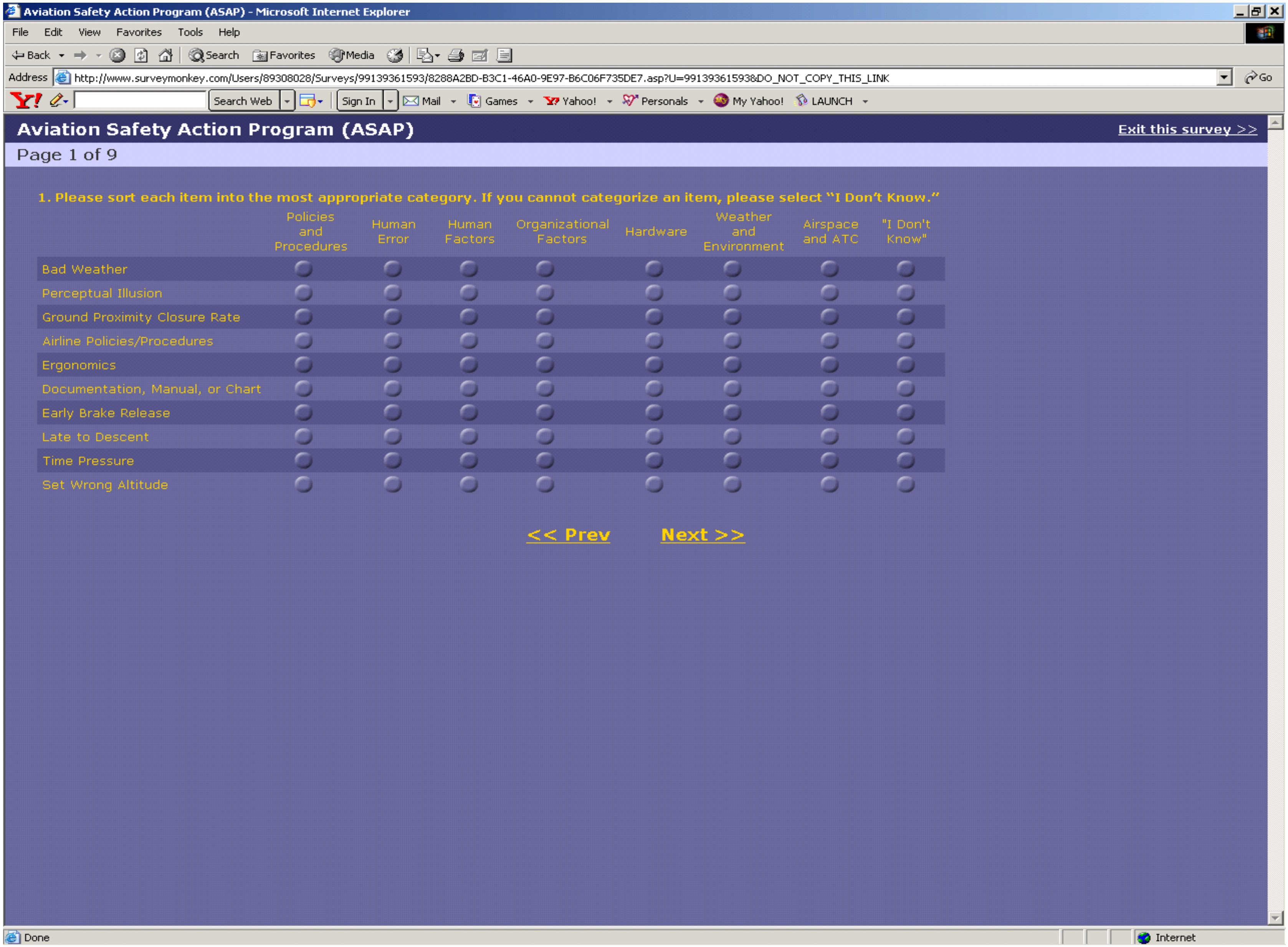
ORGANIZATIONAL FACTORS affect all the pilots at an airline. Examples include: the airline's Human Resources systems, management oversight, the allocation of organizational resources, the airline's safety culture, and the like.

HARDWARE issues concern how the airplane was designed, manufactured, and maintained. Examples include: equipment limitations, equipment malfunctions, poor design, and the like.

WEATHER & ENVIRONMENT are threats to safety that occur outside the airplane. Examples include: storm activity, difficult terrain, and the like.

AIRSPACE & ATC refer to characteristics of the National Airspace System (NAS) that can support or detract from safety. Examples include: Traffic Collision Avoidance System (TCAS) warnings, airfield layout, runway approaches, ATC requests, and radio traffic.

<< Prev   Next >>



Aviation Safety Action Program (ASAP) - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.surveymonkey.com/Users/89308028/Surveys/99139361593/8288A2BD-B3C1-46A0-9E97-B6C06F735DE7.asp?U=99139361593&DO_NOT_COPY_THIS_LINK

Aviation Safety Action Program (ASAP) Exit this survey >>

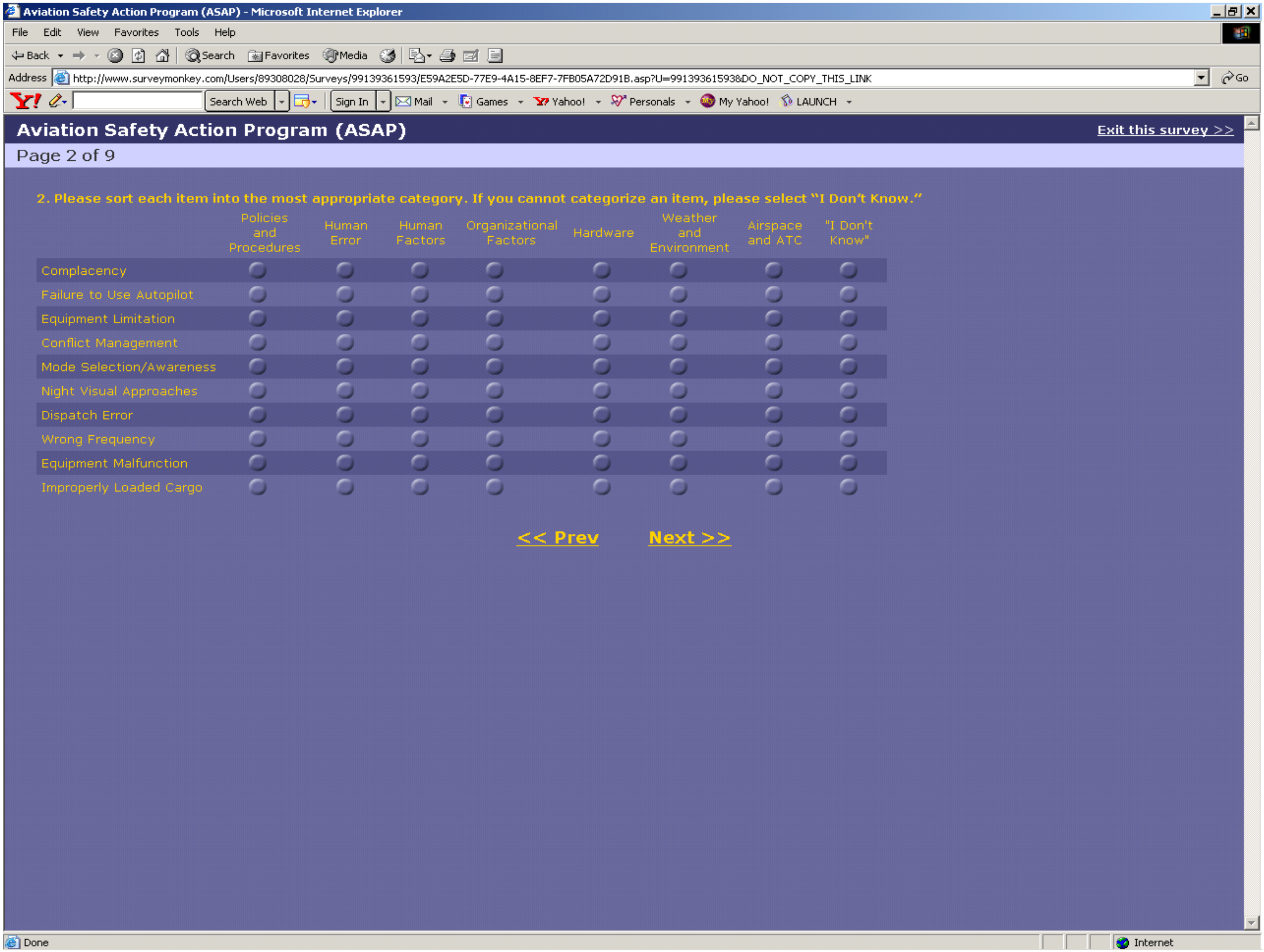
Page 1 of 9

1. Please sort each item into the most appropriate category. If you cannot categorize an item, please select "I Don't Know."

	Policies and Procedures	Human Error	Human Factors	Organizational Factors	Hardware	Weather and Environment	Airspace and ATC	"I Don't Know"
Bad Weather								
Perceptual Illusion								
Ground Proximity Closure Rate								
Airline Policies/Procedures								
Ergonomics								
Documentation, Manual, or Chart								
Early Brake Release								
Late to Descent								
Time Pressure								
Set Wrong Altitude								

<< Prev Next >>

Done Internet



Aviation Safety Action Program (ASAP) - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.surveymonkey.com/Users/89308028/Surveys/99139361593/E59A2E5D-77E9-4A15-8EF7-7FB05A72D91B.asp?U=99139361593&DO_NOT_COPY_THIS_LINK

Aviation Safety Action Program (ASAP) Exit this survey >>

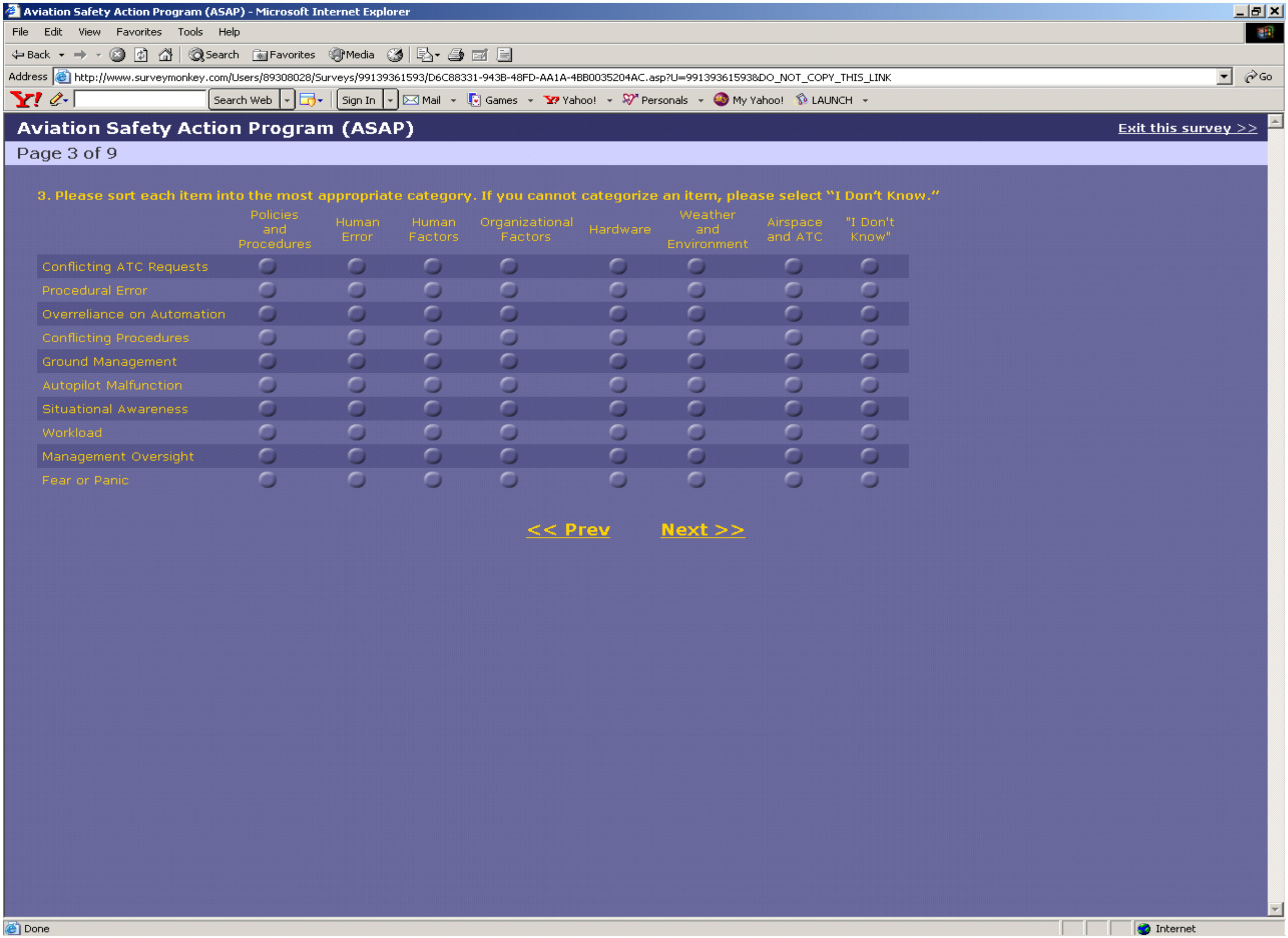
Page 2 of 9

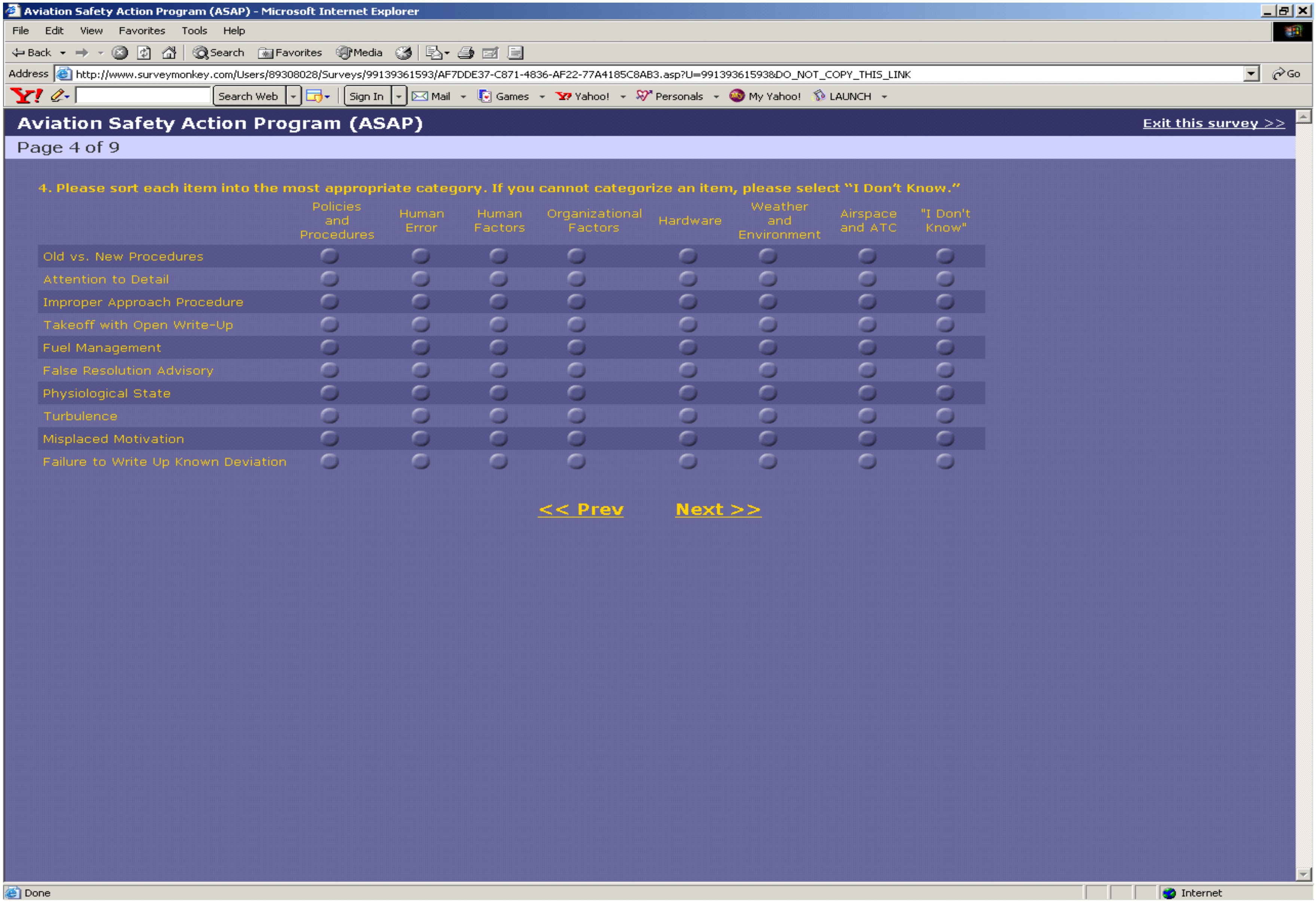
2. Please sort each item into the most appropriate category. If you cannot categorize an item, please select "I Don't Know."

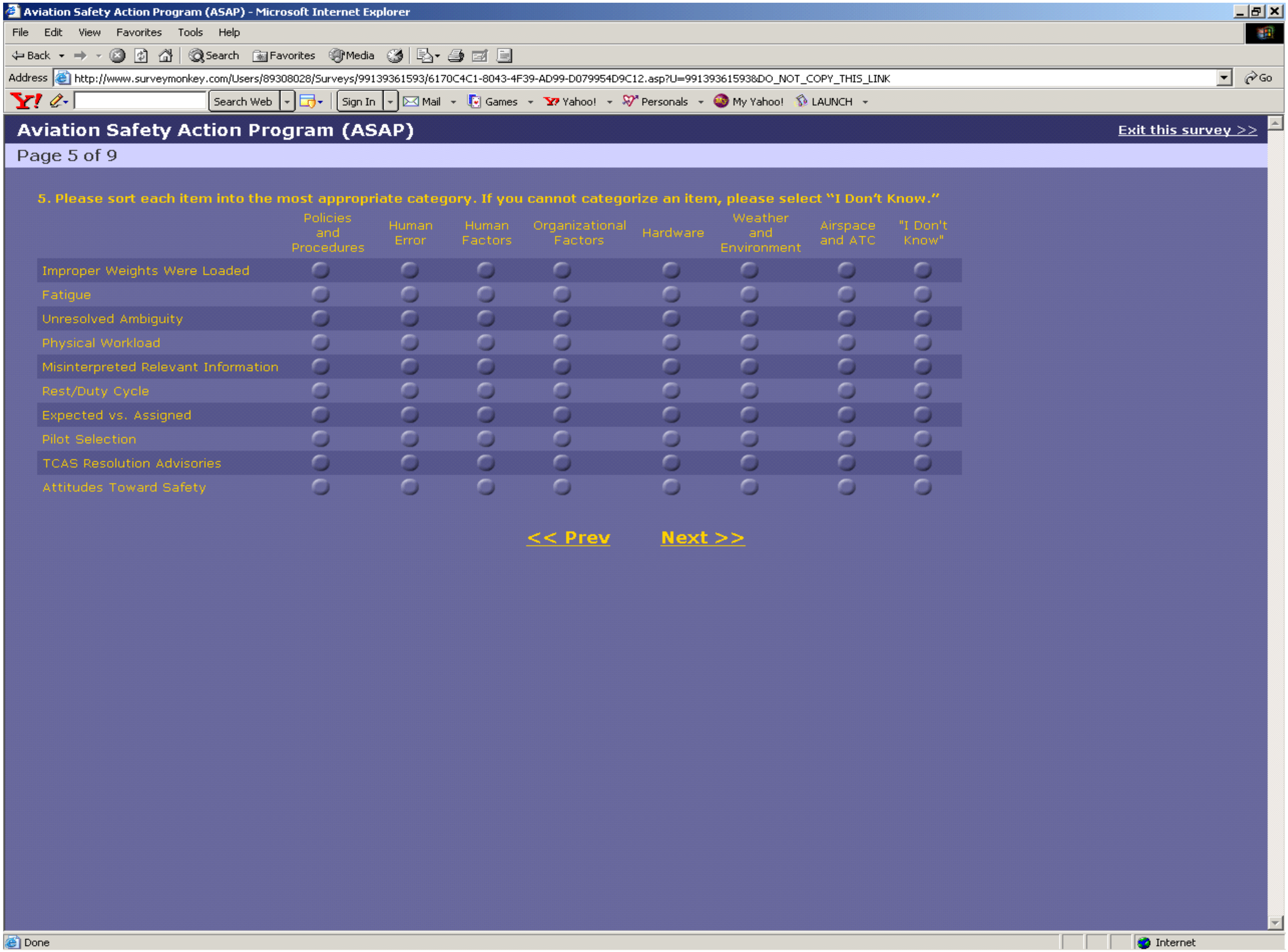
	Policies and Procedures	Human Error	Human Factors	Organizational Factors	Hardware	Weather and Environment	Airspace and ATC	"I Don't Know"
Complacency	○	○	○	○	○	○	○	○
Failure to Use Autopilot	○	○	○	○	○	○	○	○
Equipment Limitation	○	○	○	○	○	○	○	○
Conflict Management	○	○	○	○	○	○	○	○
Mode Selection/Awareness	○	○	○	○	○	○	○	○
Night Visual Approaches	○	○	○	○	○	○	○	○
Dispatch Error	○	○	○	○	○	○	○	○
Wrong Frequency	○	○	○	○	○	○	○	○
Equipment Malfunction	○	○	○	○	○	○	○	○
Improperly Loaded Cargo	○	○	○	○	○	○	○	○

<< Prev Next >>

Done Internet


Aviation Safety Action Program (ASAP) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.surveymonkey.com/Users/89308028/Surveys/99139361593/D6C88331-943B-48FD-AA1A-4BB0035204AC.asp?U=99139361593&DO_NOT_COPY_THIS_LINK
Go
Search Web Sign In Mail Games Yahoo! Personals My Yahoo! LAUNCH
Aviation Safety Action Program (ASAP)
Exit this survey >>
Page 3 of 9
3. Please sort each item into the most appropriate category. If you cannot categorize an item, please select "I Don't Know."
Policies and Procedures
Human Error
Human Factors
Organizational Factors
Hardware
Weather and Environment
Airspace and ATC
"I Don't Know"
Conflicting ATC Requests
Procedural Error
Overreliance on Automation
Conflicting Procedures
Ground Management
Autopilot Malfunction
Situational Awareness
Workload
Management Oversight
Fear or Panic
<< Prev
Next >>
Done
Internet


Aviation Safety Action Program (ASAP) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.surveymonkey.com/Users/89308028/Surveys/99139361593/AF7DDE37-C871-4836-AF22-77A4185C8AB3.asp?U=99139361593&DO_NOT_COPY_THIS_LINK
Go
Search Web Sign In Mail Games Yahoo! Personals My Yahoo! LAUNCH
Aviation Safety Action Program (ASAP)
Exit this survey >>
Page 4 of 9
4. Please sort each item into the most appropriate category. If you cannot categorize an item, please select "I Don't Know."
Policies and Procedures
Human Error
Human Factors
Organizational Factors
Hardware
Weather and Environment
Airspace and ATC
"I Don't Know"
Old vs. New Procedures
Attention to Detail
Improper Approach Procedure
Takeoff with Open Write-Up
Fuel Management
False Resolution Advisory
Physiological State
Turbulence
Misplaced Motivation
Failure to Write Up Known Deviation
<< Prev
Next >>
Done
Internet



Aviation Safety Action Program (ASAP) - Microsoft Internet Explorer

Address http://www.surveymonkey.com/Users/89308028/Surveys/99139361593/6170C4C1-8043-4F39-AD99-D079954D9C12.asp?U=991393615938DO_NOT_COPY_THIS_LINK

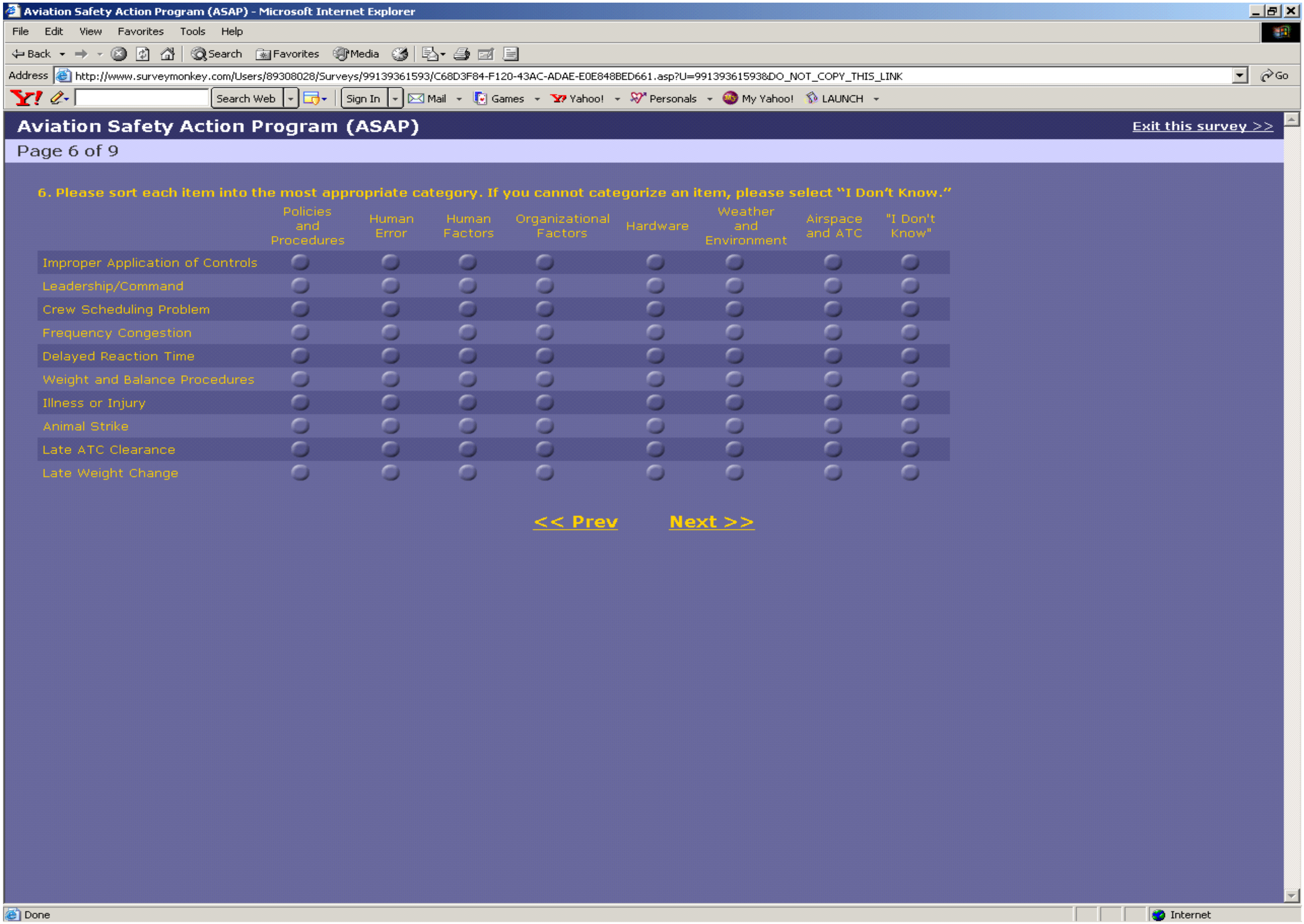
Aviation Safety Action Program (ASAP) Exit this survey >>

Page 5 of 9

5. Please sort each item into the most appropriate category. If you cannot categorize an item, please select "I Don't Know."

	Policies and Procedures	Human Error	Human Factors	Organizational Factors	Hardware	Weather and Environment	Airspace and ATC	"I Don't Know"
Improper Weights Were Loaded	○	○	○	○	○	○	○	○
Fatigue	○	○	○	○	○	○	○	○
Unresolved Ambiguity	○	○	○	○	○	○	○	○
Physical Workload	○	○	○	○	○	○	○	○
Misinterpreted Relevant Information	○	○	○	○	○	○	○	○
Rest/Duty Cycle	○	○	○	○	○	○	○	○
Expected vs. Assigned	○	○	○	○	○	○	○	○
Pilot Selection	○	○	○	○	○	○	○	○
TCAS Resolution Advisories	○	○	○	○	○	○	○	○
Attitudes Toward Safety	○	○	○	○	○	○	○	○

<< Prev   Next >>



Aviation Safety Action Program (ASAP)

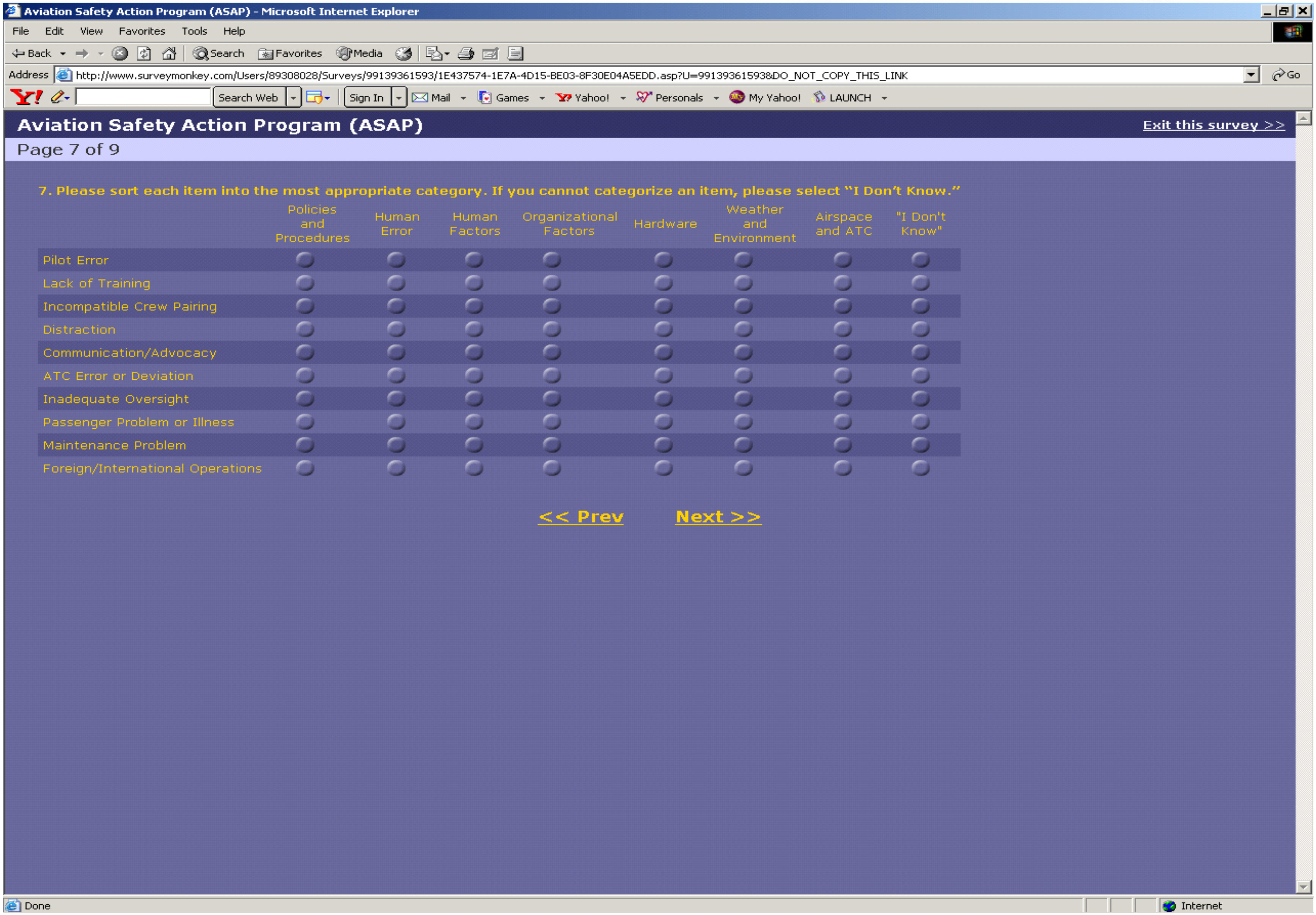
Exit this survey >>

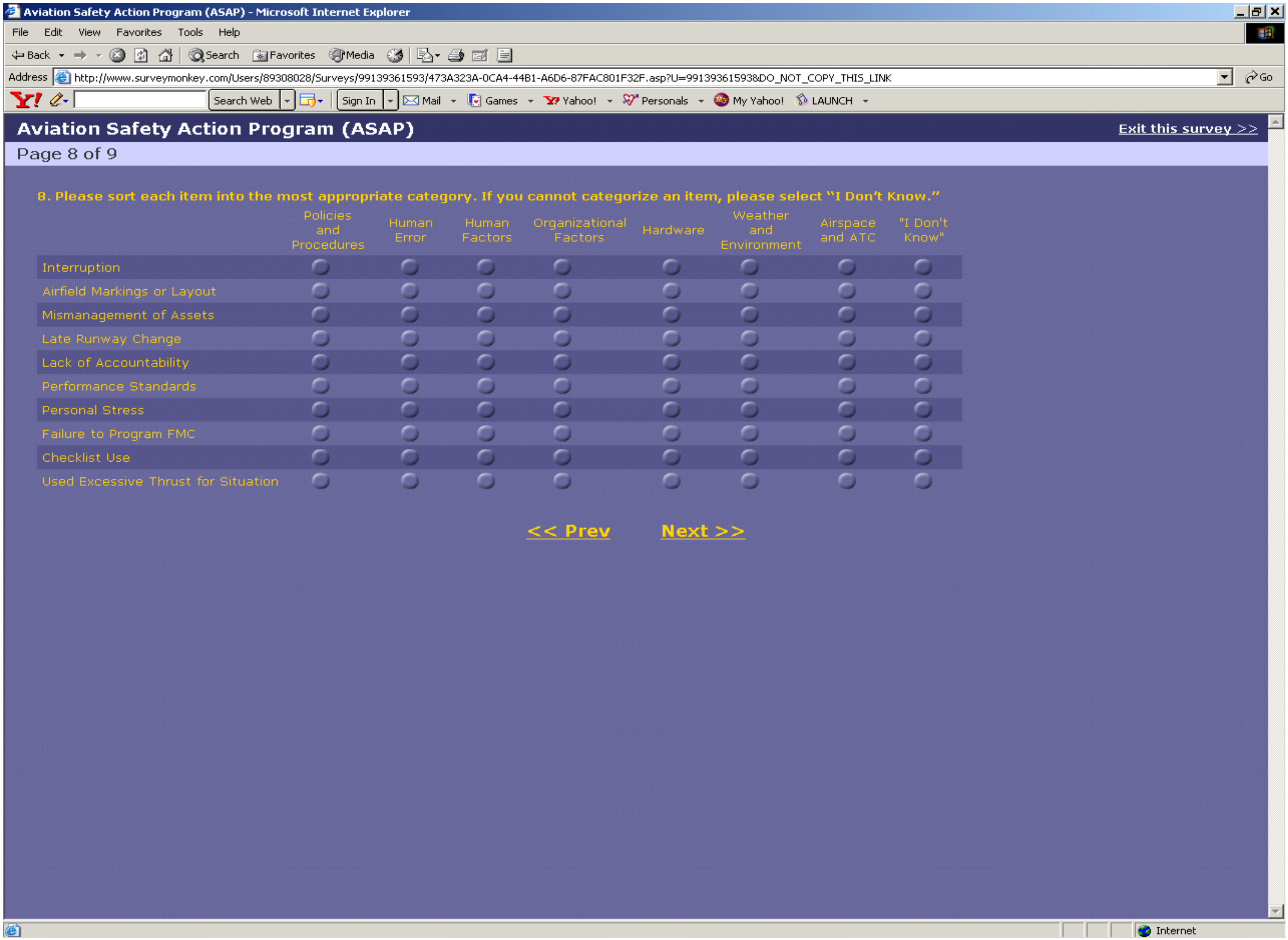
Page 6 of 9

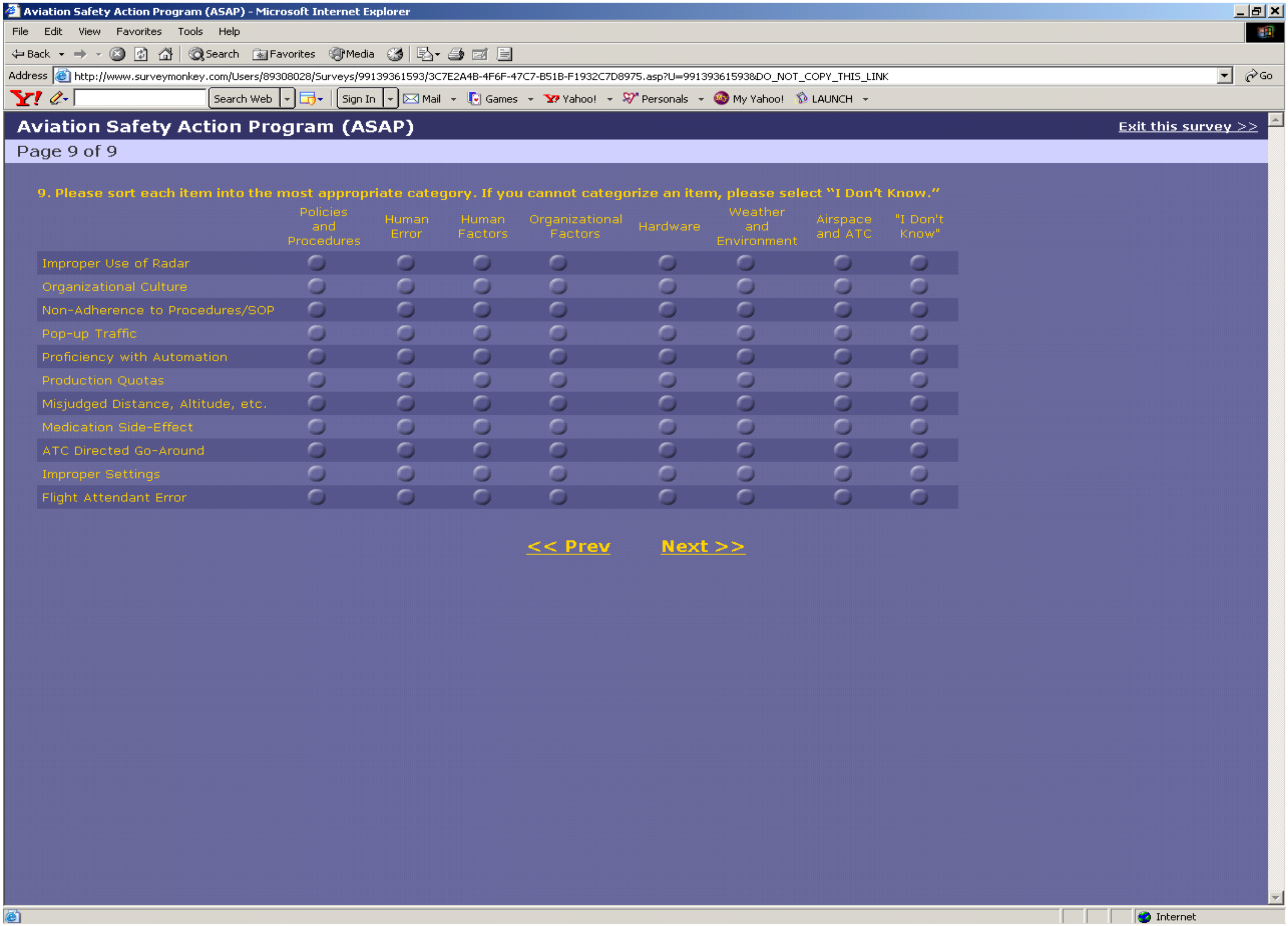
6. Please sort each item into the most appropriate category. If you cannot categorize an item, please select "I Don't Know."

	Policies and Procedures	Human Error	Human Factors	Organizational Factors	Hardware	Weather and Environment	Airspace and ATC	"I Don't Know"
Improper Application of Controls								
Leadership/Command								
Crew Scheduling Problem								
Frequency Congestion								
Delayed Reaction Time								
Weight and Balance Procedures								
Illness or Injury								
Animal Strike								
Late ATC Clearance								
Late Weight Change								

<< Prev Next >>


Aviation Safety Action Program (ASAP) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.surveymonkey.com/Users/89308028/Surveys/99139361593/1E437574-1E7A-4D15-BE03-8F30E04A5EDD.asp?U=99139361593&DO_NOT_COPY_THIS_LINK
Go
Search Web Sign In Mail Games Yahoo! Personals My Yahoo! LAUNCH
Aviation Safety Action Program (ASAP)
Exit this survey >>
Page 7 of 9
7. Please sort each item into the most appropriate category. If you cannot categorize an item, please select "I Don't Know."
Policies and Procedures
Human Error
Human Factors
Organizational Factors
Hardware
Weather and Environment
Airspace and ATC
"I Don't Know"
Pilot Error
Lack of Training
Incompatible Crew Pairing
Distraction
Communication/Advocacy
ATC Error or Deviation
Inadequate Oversight
Passenger Problem or Illness
Maintenance Problem
Foreign/International Operations
<< Prev
Next >>
Done
Internet


Aviation Safety Action Program (ASAP) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.surveymonkey.com/Users/89308028/Surveys/99139361593/473A323A-0CA4-44B1-A6D6-87FAC801F32F.asp?U=99139361593&DO_NOT_COPY_THIS_LINK
Aviation Safety Action Program (ASAP)
Exit this survey >>
Page 8 of 9
8. Please sort each item into the most appropriate category. If you cannot categorize an item, please select "I Don't Know."
Policies and Procedures
Human Error
Human Factors
Organizational Factors
Hardware
Weather and Environment
Airspace and ATC
"I Don't Know"
Interruption
Airfield Markings or Layout
Mismanagement of Assets
Late Runway Change
Lack of Accountability
Performance Standards
Personal Stress
Failure to Program FMC
Checklist Use
Used Excessive Thrust for Situation
<< Prev
Next >>
Internet



Aviation Safety Action Program (ASAP) - Microsoft Internet Explorer

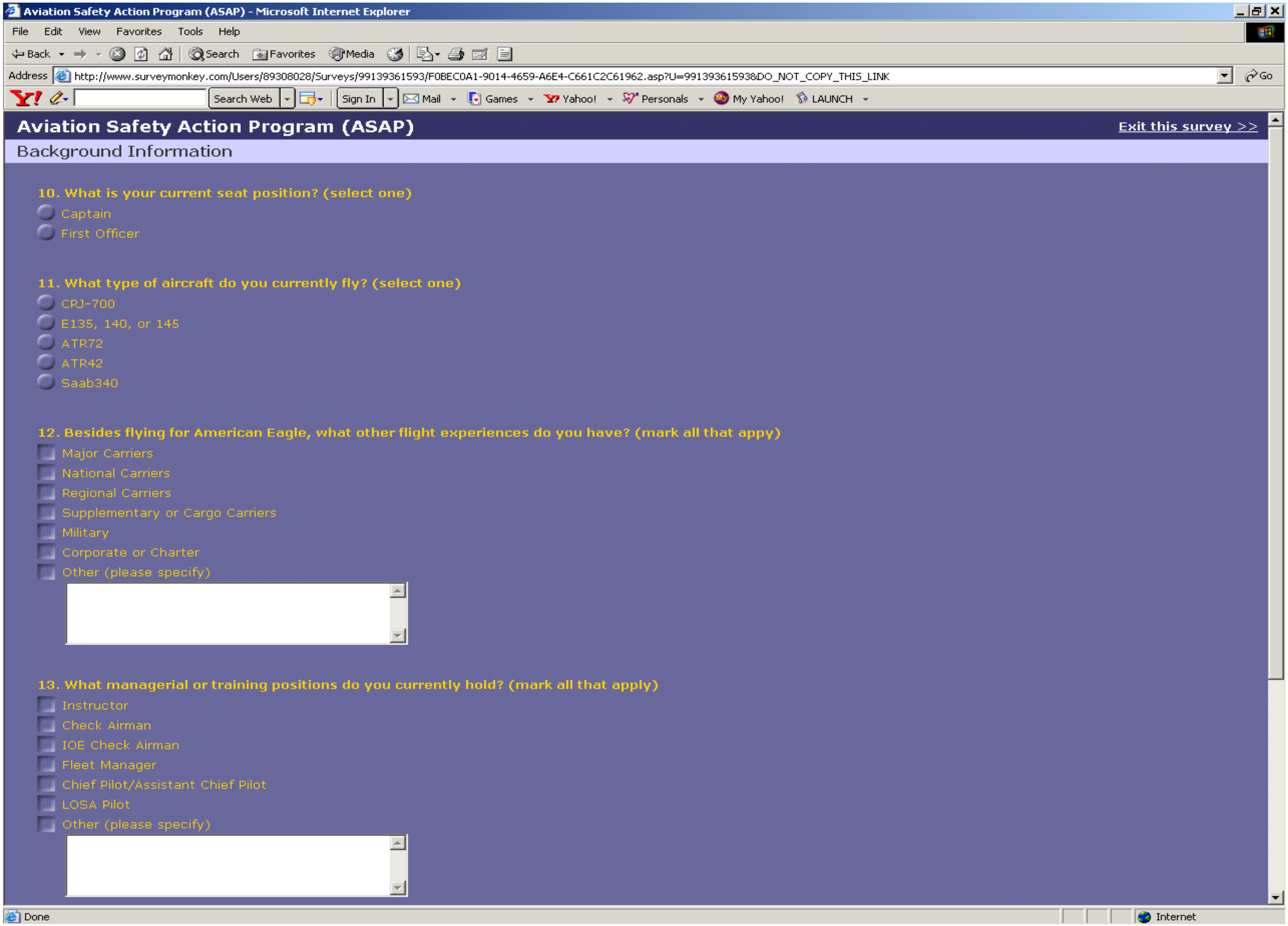
Aviation Safety Action Program (ASAP) Exit this survey >>

Page 9 of 9

9. Please sort each item into the most appropriate category. If you cannot categorize an item, please select "I Don't Know."

	Policies and Procedures	Human Error	Human Factors	Organizational Factors	Hardware	Weather and Environment	Airspace and ATC	"I Don't Know"
Improper Use of Radar								
Organizational Culture								
Non-Adherence to Procedures/SOP								
Pop-up Traffic								
Proficiency with Automation								
Production Quotas								
Misjudged Distance, Altitude, etc.								
Medication Side-Effect								
ATC Directed Go-Around								
Improper Settings								
Flight Attendant Error								

<< Prev Next >>



Aviation Safety Action Program (ASAP) - Microsoft Internet Explorer

Address http://www.surveymonkey.com/Users/89308028/Surveys/99139361593/F0BEC0A1-9014-4659-A6E4-C661C2C61962.asp?U=99139361593&DO_NOT_COPY_THIS_LINK

Aviation Safety Action Program (ASAP)

Exit this survey >>

Background Information

10. What is your current seat position? (select one)

- Captain
- First Officer

11. What type of aircraft do you currently fly? (select one)

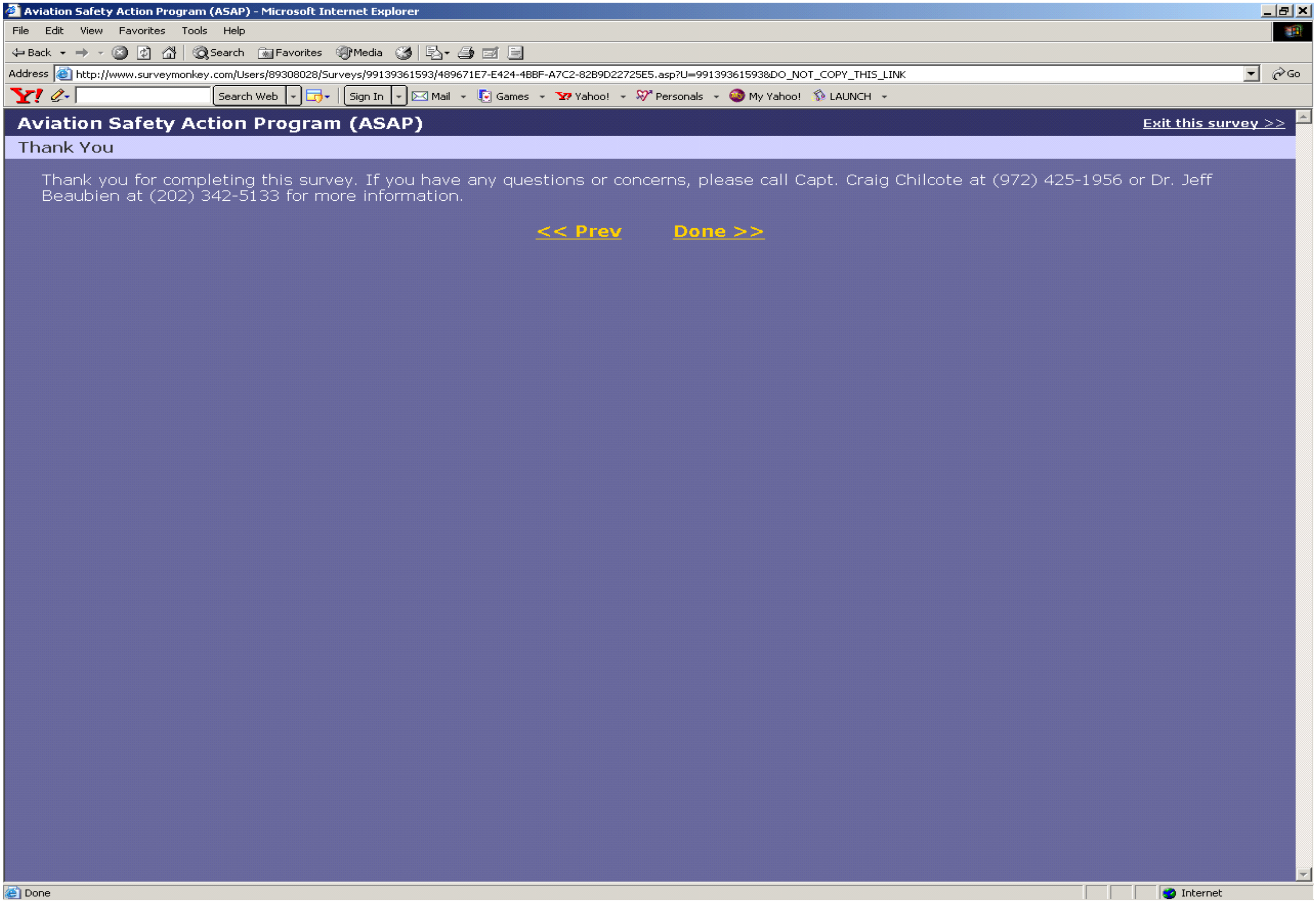
- CRJ-700
- E135, 140, or 145
- ATR72
- ATR42
- Saab340

12. Besides flying for American Eagle, what other flight experiences do you have? (mark all that appy)

- Major Carriers
- National Carriers
- Regional Carriers
- Supplementary or Cargo Carriers
- Military
- Corporate or Charter
- Other (please specify)

13. What managerial or training positions do you currently hold? (mark all that apply)

- Instructor
- Check Airman
- IOE Check Airman
- Fleet Manager
- Chief Pilot/Assistant Chief Pilot
- LOSA Pilot
- Other (please specify)


Aviation Safety Action Program (ASAP) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.surveymonkey.com/Users/89308028/Surveys/99139361593/489671E7-E424-4BBF-A7C2-82B9D22725E5.asp?U=99139361593&DO_NOT_COPY_THIS_LINK
Go
Search Web Sign In Mail Games Yahoo! Personals My Yahoo! LAUNCH
Aviation Safety Action Program (ASAP)
Exit this survey >>
Thank You
Thank you for completing this survey. If you have any questions or concerns, please call Capt. Craig Chilcote at (972) 425-1956 or Dr. Jeff Beaubien at (202) 342-5133 for more information.
<< Prev
Done >>
Done
Internet

APPENDIX B: ACCERS

AMERICAN INSTITUTES FOR RESEARCH®

Aviation Causal Contributors for Event Reporting Systems (ACCERS)
Version 1.0 Category Definitions

Policies or Procedures (1) includes Standard Operating Procedures (SOPs), charts, maps, checklists, and the like that ensure that flight operations are performed in accordance with FAA and airline regulations. Factors in this category include nonexistent, confusing, conflicting, or inaccurate policies or procedures, and the like.

Human Error (2) refers to actions that do not achieve their intended consequences, or failure to act. For example, if you incorrectly program a waypoint into the Flight Management Computer (FMC), you have made an error. Other examples include weight and balance errors, the misapplication of flight controls, inadvertent or intentional disregard of policy or procedure, improper use of the autopilot/FMS/FMC/Navigation Equipment, and the like.

Human Factors (3) are physical, psychological, and social factors that influence human performance. Examples that are likely to hinder performance include poor teamwork skills, task interruptions and distractions, complacency, stress, fatigue, task saturation, and the like. Human factors are often the precursors of errors.

Organizational Factors (4) are airline specific factors that affect all the pilots at the airline. Examples include the airline's safety culture, its rest/duty/flight times, flight or ground management, on-time performance pressures, and the like.

Hardware (5) issues concern how the airplane or ground equipment was designed, manufactured, or maintained. Examples include equipment limitations, equipment malfunctions, poor design, and the like.

Weather or Environment (6) includes threats to safety that occur outside the airplane. Examples include storm activity, turbulence, extreme temperatures, animal/bird strike, and the like.

Airspace or ATC (7) refers to characteristics of the National Airspace System (NAS) that can support or detract from safety. Examples include late, unclear, or conflicting clearances, airfield markings or layout, hear back/read back, frequency congestion, and the like.

ACCERS Version 1.0 Factor Definitions by Category Assignment

Category	***Factor***	***Factor Definition***
Policies or Procedures (1)	a. Conflicting Policies or Procedures	Two or more policies or procedures provide incompatible information or guidance regarding the same operation
	b. Confusing Policy or Procedure	Policy or procedure provides information or guidance that is difficult to understand or follow
	c. Inaccurate Policy or Procedure	Policy or procedure provides incorrect or inaccurate guidance or information
	d. International Policies or Procedures	Significant difference between International policy or procedure and Domestic policy or procedure for the same operation
	e. Lack of Policy or Procedure	No policy or procedure exists regarding this operation
	f. Old vs. New Policy or Procedure	Significant difference between old policy or procedure and new policy or procedure for the same operation
	g. Other/Please provide ___________.	The causal factor that contributed to the incident is not provided in the taxonomy. Please enter a brief description of the factor that contributed: ____________________________________.
Human Error (2)[1]	a. Altitude Planning/Awareness	Lack of awareness of altitude or poor planning with regard to establishing or maintaining proper altitude
	b. Improper Use of Autopilot/FMS/FMC/Navigational Equipment	Equipment used at the wrong time, by the wrong person, by inputting improper commands, or by interpreting output data incorrectly
	c. Inadvertent or Intentional Disregard for Policy or Procedure	Accidental or purposeful disregard of the proper policy or procedure for the operation
	d. MEL/AML Error	Failure to review MEL/AML, failure to follow MEL/AML procedure/restrictions, or

[1] For the most accurate classification within the *Human Error* Category, AIR recommends that the pilot submitting the report be queried regarding the source of the error (e.g., self/other, and if other, pilot, flight attendant, dispatcher, ATC, maintenance, or field service). Additionally, research suggests that the ACCERS taxonomy may be most effectively applied to ASAP reports when reporting pilots are queried about incident outcomes (via the *Human Error* Category and its Factors) prior to being asked to identify why the incident occurred (via the other six Categories and their associated Factors).

		following wrong MEL/AML
	e. Misapplication of Flight Controls	Aircraft controls applied at the wrong time, by the wrong person, or by applying too little/too much directional or physical control
	f. Poor Aircraft Performance Planning	Failure to consult the performance book, consulting the wrong performance book, selecting the wrong performance standard, or failing to properly apply the appropriate performance standard for a given flight condition
	g. Weight and Balance Error	Load was not distributed according to policy or procedure.
	h. Other/Please provide ____________.	The causal factor that contributed to the incident is not provided in the taxonomy. Please enter a brief description of the factor that contributed: _____________________________________.
Human Factors (3)	a. Attention to Detail	Lack of effort towards discovering the small elements of the situation or too much focus on the small elements of the situation
	b. Complacency	Lack of awareness of actual dangers or deficiencies
	c. CRM-Communication	Appropriate communication protocol was not followed
	d. CRM-Conflict Management	Conflict between crewmembers was not managed appropriately
	e. CRM-Incompatible Crew Pairing	Crewmembers do not work well together
	f. CRM-Leadership and Command	Leader failed to provide sound guidance and clear priorities
	g. Experience Level	Crewmember has too little/too much experience in this area and this had a negative impact on the situation
	h. Fatigue	Crewmember tired
	i. Fear or Panic	Crewmember afraid or suddenly frightened
	j. High Workload/Task Saturation	Amount of work to be performed in a given time period excessive
	k. Illness or Injury / Physical Health	Crewmember sick or hurt
	l. Interruption/Distraction	Interruption caused focus of attention to be shifted or diverted away from task
	m. Medication Side-Effect	Physical, emotional, or cognitive reaction to medication
	n. Motivation	Internally or externally generated drive or influence is lacking or misdirected
	o. Peer Pressure	Constraints or burdens generated by one’s friends or coworkers

	p. Personal Attitude(s) Toward Safety	Individual beliefs and feelings about safety
	q. Personal Stress	Individual mental or physical tension
	r. Proficiency/Over Reliance on Automation	Excessive dependence on mechanical aids to flight
	s. Self-Induced Time Pressure	Individual personal desire to depart the gate/arrive at the destination at the scheduled time
	t. Situational Awareness	Not knowing what is going on around you or which parameters are important, or lack of effort towards discovering important parameters
	u. Other/Please provide ___________.	The causal factor that contributed to the incident is not provided in the taxonomy. Please enter a brief description of the factor that contributed: ___________________________________.
Organizational Factors (4)	a. Airline's Safety Culture	The airline's formal and informal expectations regarding safety
	b. Employee Staffing	Failure to provide adequate staffing for particular operation
	c. Inadequate Flight Management	Improper oversight or lack of oversight from Flight Management
	d. Inadequate Ground Management	Improper oversight or lack of oversight form Ground Management
	e. Inadequate Training	Airline sponsored training regarding this operation deficient in quality, quantity, or timeliness
	f. Lack of Accountability	Chain of command unclear; no person or entity is responsible
	g. Mismanagement of Assets	Inappropriate allocation of resources
	h. On Time Performance Pressure	Pressure to ensure that the aircraft leaves the gate and arrives at its destination on time
	i. Pilot Selection	Pairing of crewmembers that are not suitable together
	j. Rest/Duty/Flight Times	Work schedule allows too little rest, or schedule frequently changed at the last minute
	k. Other/Please provide ___________.	The causal factor that contributed to the incident is not provided in the taxonomy. Please enter a brief description of the factor that contributed: ___________________________________.

Hardware (5)	a. Aircraft Damage	Physical damage to plane
	b. Aircraft Equipment Malfunction – Unable to Duplicate	Equipment fails to operate correctly when in service but maintenance personnel cannot recreate the failure
	c. Aircraft Equipment Malfunction – Query whether first time or repeated.	First known failure of equipment to operate correctly, or equipment fails again and again to operate correctly, despite check by maintenance personnel
	d. Deferred Item /MELD Equipment	Inoperative equipment
	e. Equipment Limitation	Equipment lacks the capability to perform certain needed functions
	f. Ergonomics/Poor Product Design, Placement, or User Interface	Equipment is planned, built, and/or installed in such a way that it is difficult to use
	g. Ground Equipment Inoperative or malfunctioned.	Failure of ground equipment to operate correctly
	h. Other/Please provide ____________.	The causal factor that contributed to the incident is not provided in the taxonomy. Please enter a brief description of the factor that contributed: _____________________________________.
Weather or Environment (6)	a. Animal / Bird Strike	A collision between the aircraft and an animal
	b. Excessive Cold	Atmospheric temperature very cold
	c. Excessive Heat	Atmospheric temperature very hot
	d. Icing	Considerable icing on the aircraft/runway
	e. Low Visibility/ Low Ceiling	Limited distance that one can see clearly with the naked eye
	f. Nighttime Operations	Operation was initiated, took place, or was completed in the dark
	g. Perceptual Illusion / Spatial Disorientation	Misinterpretation of visual or spatial cues
	h. Thunderstorm / Severe Weather	Extreme meteorological conditions
	i. Turbulence	Irregular motion of aircraft created by up-and-down currents of air
	j. Wind Shear	An extreme shift in wind speed or direction that occurs over a very short distance
	k. Other/Please provide ____________.	The causal factor that contributed to the incident is not provided in the taxonomy. Please enter a brief description of the factor that contributed:

		____________________________________.
Airspace or ATC (7)	a. Air Traffic Congestion	High density air traffic
	b. Airfield Markings or Layout	Lack of or poor airport markings or signs
	c. Conflicting ATC Clearance	Two or more ATC clearances provide incompatible information or guidance regarding the same operation
	d. Expected vs. Assigned ATC Clearance	ATC clearance received was different than what was anticipated
	e. Frequency Congestion	Heavy voice traffic on radio frequency
	f. Hear Back/Read Back	Failure to hear or read back the proper ATC clearance
	g. Incorrect ATC Clearance	ATC provided inaccurate guidance or information that would have resulted in a breach of safety, policy, or procedure
	h. Late ATC Clearance	ATC did not provide guidance or information in a timely manner
	i. TCAS RA/TA	A TA or RA occurred
	j. Unclear ATC Clearance	Communication followed proper protocol but guidance provided by ATC does not make sense.
	k. Other/Please provide ___________.	The causal factor that contributed to the incident is not provided in the taxonomy. Please enter a brief description of the factor that contributed: ____________________________________.