

## Comparing Evaluator Expectations and Policy Capturing Results

Jeffrey T. Hansberger, Robert W. Holt  
George Mason University

### ABSTRACT

Instructor/Evaluators (I/Es) must accurately understand the structure and content of the evaluation tool to evaluate pilots reliably and validly. This study examines how accurately I/E expectations about their evaluations correspond to their actual evaluations. Discrepancies were found between the I/Es' expectations and their actual evaluations. Estimates of grade distributions significantly deviated from actual grade distributions. Estimates of unique connections between observable behavior ratings and task ratings deviated from actual multiple regression results. Several types of errors and biases were described as the potential cause for these discrepancies.

### INTRODUCTION

Aviation safety is the top priority for many airlines and government agencies. The Instructor/Evaluator (I/E) plays a critical role in ensuring safety. The I/E's job of critically and accurately assessing pilot performance is instrumental in assuring the pilots possess adequate skills to fly safely. This study inspects the evaluation process the I/Es are instructed to follow, how well they actually follow through with this process, and their expectations of their evaluations.

The primary role of the I/E is to accurately evaluate pilot performance. One of their tasks is to detect flight performance below acceptable limits. The detection of poor performance provides them information to give accurate and helpful feedback to the pilots. The detection of unacceptable performance by the I/E also allows them to administer additional training until the skill reaches a satisfactory level. Both feedback and additional training may improve subsequent pilot performance in a number of different training environments.

The I/Es often evaluate pilot performance in a number of various settings and environments. Different evaluation environments allow the I/Es to assess different aspects of performance. The prototypical evaluation types are usually a maneuver validation, an actual line flight observation (e.g., line check), and a simulation based evaluation (e.g., LOFT, LOE).

The simulation based evaluation or a Line Operational Evaluation (LOE) has the ability to simulate actual line events with the flexibility and control to train and evaluate pilot performance in a safe environment. The LOE involved in this research was divided into event sets that correspond roughly to the phases of flight. Each event set was made up of an

event trigger, supporting conditions, and distracters. The I/Es followed a pre-constructed script and used a worksheet specific to the LOE that aided in the evaluation of each event set.

For each event set, the worksheet was designed to facilitate a specific evaluation process. The process begins with the most specific level of evaluation items, the observable behaviors. These observable behaviors are carefully identified and validated as being central to successful performance on each specific event set. The next level of item specificity is at the task (i.e., skill, or dimension) level. A skill or task level may encompass several lower level observable behaviors. The I/Es were trained to evaluate a task item based on relevant observable behavior evaluations.

Finally, the I/Es are trained to use the evaluations for the observable behaviors and tasks to aid in the individual pilot evaluations. There is also a crew based evaluation that was designed to encompass all the lower level evaluation items (e.g., observable behaviors, tasks, and individual ratings).

The LOE worksheet had the potential to simplify what could be a relatively complex evaluation process and provide the instructors with a tool for more reliable ratings. However, these benefits are contingent on the I/Es possessing an accurate understanding of the worksheet evaluation process and correctly using it on a regular basis. To the extent there are misunderstandings, errors, or biases affecting the I/Es' perception or practice of the evaluation process, there will be inconsistencies and potential inaccuracies in the pilot evaluations.

Many errors and biases have been discovered in human decision-making and performance evaluations (e.g., Klein, Orasanu, Calderwood, and Zsombok, 1995; Schneider & Schmitt, 1992). These errors and biases may occur at different stages of the evaluation process. The evaluation process can be elaborated into two general processes or stages, a memory retrieval and a judgement process.

The memory retrieval process for an evaluation task involves the recall of stored knowledge from past experience and training. It also involves the recall of past evaluation experiences that may be used to compare the current situation to. The judgement process, however, involves how the I/E uses the retrieved and observed information. The LOE worksheet is an example of an evaluation aid to the judgement process. Each of these processes are susceptible to their own type of errors and biases among the I/Es.

Several errors and biases may affect the memory retrieval process are: the base-rate fallacy, availability heuristic, recency effect, and the representativeness heuristic. The base-rate fallacy is the tendency to be overinfluenced by distinctive or extreme cases (Nisbett, Borgida, Crandall, and Reed, 1976). A failure or an extremely excellent performance is more easily remembered than all the standard performances that are observed by the I/E. This may affect the criteria used by the I/Es to grade others or their perceived distribution of pilot performance in their fleet/airline.

The availability heuristic is the tendency to recall very salient or often used judgments, ideas, and facts from memory (Reyes, Thompson, Bower, 1980). An I/E practicing the availability heuristic may be over relying on particularly salient criteria or past experience for their evaluations.

The recency effect is when recent information, evaluations, or experiences have a larger effect on a judgement than other instances (Miller & Campbell, 1959). An example of a recency effect would be if an I/E relies on a recently observed crew performance as judgment criteria instead of the airline's standards.

Lastly, the representativeness heuristic is the tendency to make judgements based on how well they represent or match a mental prototype (Fischhoff & Bar-Hillel, 1984). This heuristic can lead an I/E to ignore other relevant or extenuating information if it differs from their prototype.

The possible errors and biases that may affect the judgement and evaluation process differ from the memory retrieval process. These potential errors and biases are: the leniency error, central tendency error, halo error, stereotyping, and the similar-to-me phenomenon (Schneider & Schmitt, 1992). The leniency error refers to the fact that some I/Es give consistently higher ratings than others. This would make it difficult to accurately compare evaluations across different I/Es.

The central tendency error occurs when the I/Es do not use the extreme scores on the rating scale. This type of error would result in lower scores for the superior performers and higher scores for the below average performing pilots.

The halo error occurs when a particularly positive or negative act or performance inaccurately biases an I/E's evaluations. An I/E who observed a pilot handle an abnormal/emergency situation well at the beginning of the flight might judge their following performance higher even if it does not deserve to be. The reverse also holds with negative actions or events.

The stereotyping effect is a type of halo effect where the judgement of a pilot's performance is partially based on the pilot's group membership instead of the pilot's actual performance. An I/E may know what

airline or military servicea pilot formerly flew at and have a negative opinion of that airline. A stereotyping effect would influence the I/Es judgements negatively for this pilot.

The similar-to-me phenomenon occurs when the I/E judges the characteristics and attitudes in another pilot relative to his or her own characteristics. The I/E may compare pilots to "what they would do" and provide the pilot with a potentially higher or lower evaluation than they deserve.

Very little research has investigated the extent errors and biases are occurring in an aviation setting among I/Es (Williams, Holt, & Boehm-Davis, 1997). Even less research has been conducted on pinpointing exactly what errors and biases are present, if any. To the extent there are errors and biases influencing I/E evaluations, there will be inaccuracies in the evaluation process and in the expectations of grade distributions among the pilot population.

There are two types of inaccuracies that may result from the described errors and biases, an inaccurate view of the LOE worksheet evaluation *process* and *outcome*. The LOE worksheet evaluation process was described earlier as the process of using the lower level evaluation items as a basis for the higher level items. The LOE worksheet evaluation outcome would be the final evaluations given and their distributions among the pilot population.

This study examined how well the I/Es' expectations of links in the LOE worksheet process matched actual results. This study also investigated to what extent, if any, the I/Es' expectations of the grade distributions were distorted from their actual grade distribution they gave the pilot population.

## METHOD

### Subjects

Eleven I/Es from one fleet in a regional airline participated in this study. This sample of I/Es represented approximately 85% of the total I/Es for this fleet. All the participating I/Es had received their required training and had evaluated LOEs in the past six months.

### Materials

Part 1 of the questionnaire evaluated how the I/Es were using the evaluation items in the LOE worksheet. Part 1 of the questionnaire contained twelve pages. Each page presented the observable behaviors and task items for an event set from their current LOE. For each event set (i.e., page), the I/Es were asked to draw a line from each observable behavior to the task items they thought was related. For example, an I/E might believe the observable behavior, "Crew briefs takeoff conditions to include turbulence" is related to the task of "Handling of departure in turbulence" and

“Addressing takeoff and departure issues”. The I/E would draw a line from the observable behavior to each of these two tasks. The I/Es completed this task for each observable behavior for every event set (i.e., 11 event sets). The LOE was designed for each successive level of evaluation to be based on the prior level (e.g., the task is dependent on relevant observable behaviors).

Part 2 of the questionnaire evaluated the I/Es’ expected distribution of the grades for their Line Checks, Maneuver Validation, individual ratings for the LOE (Pilot-in-command (PIC) & second-in-command (SIC)), technical and crew resource management (CRM) grades for the LOE, and the observable behaviors for the LOE.

For each of the above categories, they were asked to record what percentage out of 100% they “would expect to see for pilot grades over the last 6 months”. All the categories above are rated on a four-point scale except the observable behaviors for the LOE, which was rated on a three-point scale. The four-point scale ranged from “unsatisfactory” to “above standard” and the three-point scale ranged from “not observed” to “fully observed.” The pilots were asked to double-check that each row of estimates added to 100.

#### Procedure

Eleven I/Es from one fleet of the regional carrier completed the part one and part two of the questionnaire, in that order. This was done during a monthly meeting of this I/E group. Researchers were present to answer questions and resolve any ambiguities in the task or response format.

Actual LOE evaluation Data over the duration of approximately 12 months was also collected. Only the data specific to the before mentioned fleet was used in the analysis. This data was used to evaluate how accurately the I/Es perceptions of the evaluation process matched their actual evaluation process.

## RESULTS

### Evaluation Process

The evaluation process was analyzed by comparing the I/Es’ perceptions of the evaluation process to how they actually conducted their evaluations. Multiple regressions were used to determine how the I/Es were using the evaluation items during the LOE. A multiple regression was conducted for each event set in which the tasks (2 technical and 1 CRM) were regressed on the observable behaviors.

The semi-partial  $r$  was used to indicate the links between the observable behaviors and the tasks because it represents the unique variance accounted for by each observable behavior. A more liberal significant value of .10 was used to estimate the significance of

the semi-partial  $r$  scores. This was done primarily for the small sample size and to reduce type II error.

Due to space limitations, only the best and worst cases will be presented here with a summary of the overall results. The event set that represented the best overall match between the I/Es’ perceptions and their actual ratings is illustrated in Figure 1. Signal detection theory was used to categorize the results for the event sets. Table 1 illustrates the basic components of signal detection theory. Table 2 applies the signal detection theory annotation to the results of this event set. As shown in Table 2 and Figure 1, all three links empirically found were expected by the I/Es. However, only five of the nine (55.6%) non-significant relationships between observable behaviors and tasks from the data were correctly identified by the I/Es (5 correct rejections & 4 false alarms).

The worst overall event set is illustrated in Figure 2. Table 3 summarizes the results using the signal detection theory. The one empirically found link was not one of the I/Es’ expected links (miss). Six of the eleven (54.5%) non-significant empirical relationships between observable behaviors and tasks were correctly identified by the I/Es (6 correct rejections & 4 false alarms).

Table 4 illustrates the results across all eleven event sets. The overall percentage of matched links or hits for was 67.9% (19 hits & 9 misses). The overall non-significant empirical relationships between observable behaviors and tasks were correctly estimated by the I/Es 54.1% (59 correct rej. & 50 false alarms) of the time. This represents a large overestimation of expected links that were not found in the actual data. This finding of overestimating the links was present for every event set with the exception of one.

### Evaluation Outcomes

The evaluation outcomes were analyzed by comparing the I/Es’ expected distributions of grades from Part 2 of the questionnaire to their actual grade distribution of grades for the LOE. The I/Es provided distribution estimates for the observable behavior, the task, and the individual pilot evaluation level of the LOE. These estimates were given as percentages.

Due to space limitations, the overall averages for the I/Es’ expectations at only the task level are illustrated in Figure 3. Also illustrated in Figure 3 for comparison purposes, is the distribution of task grades given for the LOE. The I/Es are overestimating all the grades except for the “3” grade (“standard”), which is largely underestimated. This same trend holds at the observable behavior and individual pilot level evaluations as well.

Frequency distributions were calculated from the percentage estimates given by the I/Es at each evaluation level. Chi-square analyses were done comparing the expected distribution to the actual

distribution. Considering each evaluation as an independent event, the expected observable behavior ( $\chi^2 (2) = 5823.2, p < .01$ ), task ( $\chi^2 (3) = 2333.6, p < .01$ ), and individual pilot ( $\chi^2 (3) = 457.8, p < .01$ ) distributions were all significantly different than the actual LOE grade distributions.

## DISCUSSION

The vital role the I/E plays in aviation safety requires that they are able to accurately and reliably evaluate pilot performance. It is clear from these research results that there are differences between the I/Es' expectations and actual evaluations for both the LOE evaluation process and outcomes. These differences may represent a number of various errors and biases enacted by the I/Es for both the evaluation process and outcomes.

The results from the evaluation process showed a large overestimation bias for the number of links between observable behaviors and tasks. The I/Es thought there were many more links than the data showed they were actually using. This perception of relationships where none actually exist is called an illusory correlation (e.g., Crocker, 1981).

When the designers of the LOE items were asked informally which results most closely matched their design intentions, they chose the I/E expectations. The design intentions were to include as many relevant observable behaviors to aid in the higher-level evaluation items.

Several possible errors and biases were described in the introduction. Based on the results, many of these biases may be hypothesized to play a role in the I/E evaluations. The most likely candidates to play a role in the evaluation process are the availability heuristic and the representativeness heuristic. The I/Es are obviously depending on other sources of information besides the observable behaviors in the LOE when making their evaluations.

The availability heuristic postulates that the I/Es might be relying on a set of easily accessible criteria information that they have had a lot of experience with, which is very salient. The representativeness heuristic, however, postulates the I/Es might have their own prototypes of what they consider "Unsatisfactory" or "Standard" performance. This information might be used in place of given judgement criteria or information, even if this information is relevant to the evaluation.

Other types of errors or biases might be responsible for the differences found for the LOE outcomes. The base-rate fallacy is a strong candidate for explaining the I/E distribution discrepancy. The I/Es overestimated the extreme scores for all three evaluation levels, which suggests they might be

overinfluenced by the extreme cases in their memory recall of past evaluations. If this process occurs during the LOE assessment, it may have an effect on evaluation outcomes.

If these errors and biases are confirmed as playing an influential role in the evaluations, there are implications for training and future evaluation of rater training. The discovery of systematic errors or biases in an evaluation process can be addressed and minimized in training. However, the specific error or bias must be known by the trainers in order to address it. It is also possible using this method to evaluate the training of the I/Es and evaluate the increase or decrease of the targeted errors or biases.

This research discovered there were differences in how the I/Es perceive both the evaluation process and outcomes of an LOE. Several possible errors and biases were hypothesized to be a possible cause for the I/Es' discrepancies. The biases described here are the most likely given the current results. Other errors and biases may also play a role in the I/Es' evaluations but would require further research to confirm.

## REFERENCES

- Crocker, J. (1981). Judgment of covariation by social perceivers. Psychological Bulletin, *90*, 272-292.
- Fischhoff, B., & Bar-Hillel, M. (1984). Diagnosticity and the base rate effect. Memory and Cognition, *12*, 402-410.
- Klein, G.A., Orasanu, J., Calderwood, R., & Zsombok, C. (eds.) (1995). Decision making in action: Models and methods. Norwood, NJ: ABLEX Publishing.
- Miller, N., & Campbell, D. (1959). Recency and primacy in persuasion as a function of the timing of speeches and measurements. Journal of Abnormal and Social Psychology, *59*, 1-9.
- Nisbett, R. Borgida, E., Crandall, R., & Reed, H. (1976). Popular induction: Information is not necessarily informative. In J.S. Carroll & J. Payne (Eds.), Cognition and Social Behavior. Hillsdale, N.J.: Erlbaum.
- Schneider, B. & Schmitt, N. (1992). Staffing Organizations. Prospect Heights, IL: Waveland Press.
- Reyes, R., Thompson, W., & Bower, G. (1980). Judgmental biases resulting from differing availabilities of arguments. Journal of Personality and Social Psychology, *39*, 2-12.
- Wickens, C.D. (1992). Engineering Psychology and Human Performance. HarperCollins Publishers.
- Williams, D.M., Holt, R.W., and Boehm-Davis, D.A. (1997) Training for inter-rater reliability: Baselines and benchmarks. In Proceedings of the Ninth International Symposium on Aviation Psychology.

Response	State of the World		
		<b>Signal</b>	<b>Noise (no signal)</b>
	<b>Yes</b>	Hit	<i>False alarm (FA)</i>
<b>No</b>	<i>Miss</i>	Correct rejection (CR)	

Table 1. The four outcomes of signal detection theory (Wickens, 1992). Italics represent an incorrect response.

Expected LOE Results	Actual LOE Results		
		<b>Link</b>	<b>No Link</b>
	<b>Link</b>	3 hits	<i>4 FAs</i>
<b>No Link</b>	<i>0 misses</i>	5 CRs	

Table 2. Signal detection results for the BEST overall match of expectations to actual evaluations for an event set (italics represent incorrect responses, see Table 1).

Expected LOE Results	Actual LOE Results		
		<b>Link</b>	<b>Noise</b>
	<b>Link</b>	0 hits	<i>4 FAs</i>
<b>No Link</b>	<i>1 misses</i>	6 CRs	

Table 3. Signal detection results for the WORST overall match of expectations to actual evaluations for an event set (italics represent incorrect responses, see Table 1).

Expected LOE Results	Actual LOE Results		
		<b>Link</b>	<b>No Link</b>
	<b>Link</b>	19 hits	<i>50 FAs</i>
<b>No Link</b>	<i>9 misses</i>	59 CRs	

Table 4. Signal detection OVERALL results of for the match of expectations to actual evaluations for all event sets (italics represent incorrect responses, see Table 1).



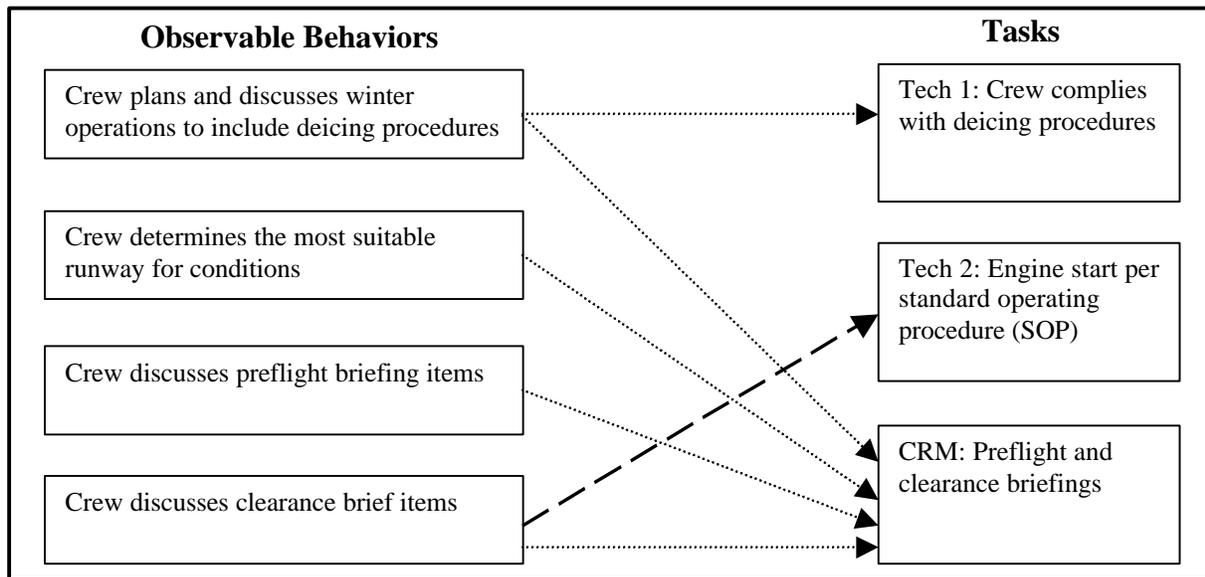


Figure 2. The observable behaviors and tasks occurring for the WORST overall match of expectations to actual evaluations for an event set. The dotted lines represent links expected by the I/Es but not found in their actual evaluations. The dashed lines represent links found in the actual data but not expected by the I/Es (There were no matches between the I/Es' expectations and actual evaluations for this event set.)

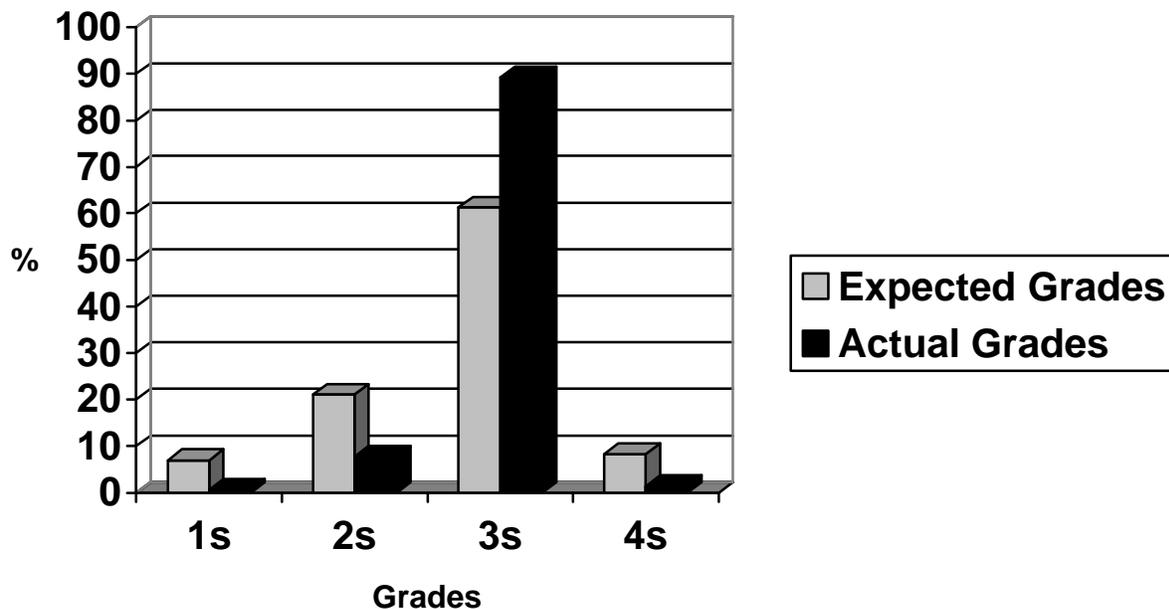


Figure 3. The distributions of the each expected grade and each actual grade given for the LOE at the task level.

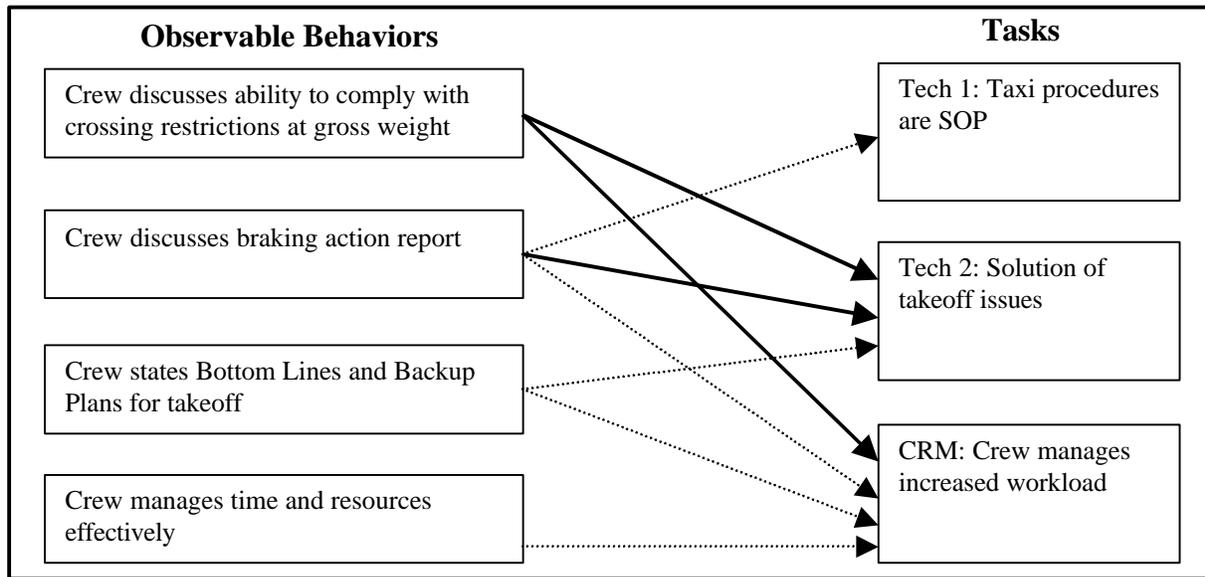


Figure 1. The observable behaviors and tasks occurring for the BEST overall match of expectations to actual evaluations for an event set. The dotted lines represent links expected by the I/Es but not found in their actual evaluations. The solid lines represent matched links from both expectations and actual evaluations (Links found in the actual data but not expected by the I/Es did not occur in this event set).