Evaluating human-machine-interfaces for making binary choices: why measuring uncertainty is important and how to do it

A. Baier and A.C. Zimmer

Engineering Psychology Unit, University of Regensburg, Germany

Abstract - Many safety-relevant tasks in control or diagnostics require binary choices such as 'conflict vs. separation' in air-traffic control, 'normal vs. pathological' when interpreting x-ray pictures, or 'permitted vs. forbidden' when inspecting airport security scans. Deciders often are uncertain, but nevertheless required to decide between two alternatives, that is, they have not only to decide upon an action, but also about the admissible level of uncertainty. If the accepted level of judgment certainty is not taken into account, the sequence of decisions does not capture the full picture of the underlying decision process. Differences in judgment certainty are relevant, because they reflect not only the adequacy of the human-machine-interface that is evaluated, but also the differences in expertise of the decider and the requirements of the actual situation or task. Therefore, capturing both judgment certainty and discrimination performance is essential. A comparison of different human-machine-interfaces (for air traffic control) is used to illustrate a methodological approach, which allows for integrated analyses of decision processes based on receiver-operator-characteristics and practical guidelines for the evaluation of human-machine-interfaces for safety-relevant operation procedures are provided.

INTRODUCTION AND THEORY

Many tasks in control or diagnostics require the deciders to make binary choices such as 'conflict vs. separation' in air traffic control, 'normal vs. pathological' when interpreting x-ray pictures, or 'permitted vs. forbidden' when inspecting airport security scans. Although deciders often are uncertain, a choice between the two alternatives that formally can be defined as 'positive' and 'negative' nevertheless is required, therefore causing a dilemma: For achieving a high performance in discriminating between positive and negative cases, the decider has to maximize the fraction of true positives (hit rate) out of the total actual positives, and minimize the fraction of false positives (false alarm rate) out of the total actual negatives at the same time. Unfortunately, in all cases in which the decider is not absolutely certain about the correctness of his decision, raising the true positive rate requires to classify more and more cases as positive even though uncertain if they are indeed positives. Hence, increasing the true positive rate is inevitably accompanied by an increased likelihood for a false alarm and vice versa. The decider has to choose between striving for the maximization of the former and therefore tolerating more false alarms, and striving for minimization the latter and therefore diminishing the number of true positives, or aiming to reach a balance between the resulting true positive and false alarm rate. The discrimination performance, however, stays unaffected from this choice, which shall be illustrated with the following example. If the decider is absolutely uncertain but wants to ensure that no positive case is missed, he classifies all cases as positives. Consequently, all negatives will be also classified as positives, resulting in a discrimination performance at chance level. The same performance results when, for instance, both half of the positives as well as half of the negatives are classified correctly. Though the discrimination performance in both examples is the same, the underlying decision process is a different one, because in the latter the decider accepts a higher degree of uncertainty. The result of the choice about how much uncertainty should be accepted is termed 'criterion' and categorized into 'liberal', 'conservative' and 'neutral' response behaviour. The liberal criterion reflects the tendency to classify uncertain cases preferably as positives rather than negatives [1].

WHY MEASURING UNCERTAINTY IS IMPORTANT

For the evaluation of human-machine-interfaces used to make binary choices, analyzing both the performance and the judgment certainty is of mayor importance, because the outcome of the decision-process in terms of true positive and false positive rates is not only a result of how well the human-machine-interface supports the decider in discriminating between positive and negative cases, but also is a result of the level of uncertainty the decider is willing or allowed to accept. There are several important factors in the context of evaluating human-machine-interfaces that, besides the design characteristics of the human-machine-interface, impact on the decision of the decider: His expertise

with the interface (i) and the task (ii), as well as the characteristics of the task (iii) and the situation (iv). Interactions between these factors can additionally impede the interpretation of the results. The following examples shall point out how they can lead to counterintuitive effects.

i) While a decider is able to discriminate between the majority of the positive and negative cases presented with the interface he or she is highly used to, a novel interface might cause a higher degree of uncertainty, encouraging him or her to apply a more liberal response criterion. Therefore, both a higher true positive and false positive rate result with the novel interface. The discrimination performance with the novel interfaces, however, could well be equal, better, or worse as with the traditional interface.

ii) The same is true when the decider possesses profound expertise with the tasks. He or she might be equally certain about the presented cases, but achieve an equal, better, or worse discrimination with the novel interface.

iii) Another possibility is, that certain tasks such as vertical distance judgments cause a higher uncertainty when, for instance, presented with a 2D compared to a 3D visualization, whereas for horizontal distance judgments the contrary might be true. Such an interaction between human-machine-interface and task-characteristics might conceal existing differences, by resulting in both an overall comparable discrimination performance and judgment certainty, though clear advantages exist for each kind of visualization dependent on the task to be achieved.

iv) In general, the risks and incentives certain situations comprehend play an important role when deciding how much uncertainty is acceptable. While an air traffic controller often has only one opportunity to decide, and a wrong decision is likely to cause fatal consequences, he or she will apply a liberal response criterion. Medical doctors or airport security officers, in contrast, might show a stronger trend towards conservative response behaviour because they face different demands. If uncertain, they might decide to conduct another test in order to re-evaluate the diagnosis before informing a patient about a radical result or allowing a passenger to enter an airplane.

These examples highlight that an objective evaluation requires separating discrimination performance from response behaviour to enable a correct interpretation of the results.

METHODS FOR EVALUATING PERFORMANCE AND UNCERTAINTY

Selecting test cases and rating procedures

As a basis for the measurement, a representative set of cases that includes as many typical task characteristics as possible has to be presented, and is so much the better the more cases are used [1]. To facilitate the interpretation of the results, it is helpful to present the decider with an equal number of positive and negative cases in a randomized order. Right after the presentation of each case, the decider is asked to classify it as positive or. To do so, a rating scale with an at least ordinal scale of measurement should be used. We recommend using a six-point-rating scale that allows for capturing an interval level of measurement, a so-called Likert-scale. The even number of response options forces the decider to indicate a tendency towards one of the two endpoints. The interval scale allows conducting a broad variety of statistical calculations on the resulting data. More than six options tend to overload the decider, whereas less might limit the decider in expressing the perceived level of certainty and the comprehensiveness of the resulting information.

Calculating hit and false alarm rates and visualizing performance: The ROC curve

After the rating procedures have been completed, first both hit and false alarm rates for each response option and human-machine-interface that shall be evaluated are calculated. Afterwards, and beginning with the resulting hit and false alarm values for the option 'certainly positive', the hit and false alarm rates of the next response option 'probably positive' and so forth are added, producing pairs of hit and false alarm values that increase with adding each option until a value of 100% results. Based on this, a so-called receiver operating characteristic (ROC) curve can be created to demonstrate the performance resulting with each human-machine-interface. To do so, the values are plotted into a coordinate system in which the ordinate represents the hit rate and the abscissa the false alarm rate, and connecting the data points including the zero scale marks.

Isolating performance from uncertainty: The area under the ROC curve

The area under the ROC curve (AUC) indicates the likelihood with which the decider detects a true positive case correctly as such when presenting a randomly chosen case out of all cases on which the ROC curve is based. The AUC value can vary between the two values 0 and 1 of which the latter indicates a perfect discrimination performance. A result of 0.5 signifies a performance at chance level. The AUC value therefore serves as a measure for expressing the discrimination performance independently from the underlying judgment certainty, since it solely depends on the size of the area under the curve and not on its shape. That is, the criterion can vary on the graph, therewith representing different response criteria that could be applied when uncertain about if the displayed case is positive or negative. The discrimination performance, however, stays the same no matter which response criterion the decider applies for each response option [1]. This facilitates an objective comparison of different human-machine-interfaces superior to comparing hit or false alarm values directly because the latter depends on the response criteria the deciders apply.

Comparing performance while controlling judgment certainty: The zROC graph

By transferring the hit and false alarm values into standardized z-values, connecting them with a straight line by calculating a linear equation, and plotting them into a coordination system with equally standardized axis, for any desired hit rate the resulting false alarm rate can be predicted and vice versa. These so called zROC graphs facilitate the evaluation of different human-machine-interfaces in a way that goes beyond comparing the discrimination performance on the basis of the AUC values. The evaluator now can choose from any criterion a decider might want to apply in order to deal with uncertainty, and compare the resulting performance between the different human-machine-interfaces. The fact that the deciders may have applied different criteria with each interface is irrelevant. Please note that determining zROC graphs is so much the better, the more response options have been given. A binary response option, however, does not allow the calculation of a zROC graph, because it only allows calculating one point of the graph and the required information for determining the slope of the zROC graph is missing unless, for instance, assumptions can be derived from similar experiments.

HOW TO GATHER, ANALYZE AND INTERPRETE YOUR DATA

Comparing expert performance with a traditional and a novel interface: An example from air traffic control

To illustrate how the results from comparing different human-machine-interfaces for making binary choices can be analysed and interpreted with the above described methodology, we use a data set from a recent study in which we compared different visualizations for air traffic controller workstations [2]. Amongst others, we used a representative set of 32 safety critical air traffic scenarios that were presented to 12 air traffic controllers whose task it was to classify each scenario as conflict (positive) or separation (negative) using a 2D visualization similar to the one used today at air traffic controller workstations as well as a stereoscopic 3D visualization. Each scenario started 45 seconds before the

respective aircraft actually collided or reached the closest point of approximation in case they missed each other, and was shown for exactly 10 seconds before it was blinded out. After each scenario, an entry mask with a six-point-rating scale and the response options ‘certainly positive’, ‘probably positive’, ‘maybe positive’, ‘maybe negative’, ‘probably negative’, and ‘certainly negative’ was presented. This allowed the air traffic controllers to express their certainty about the outcome of each scenario. Table 1 shows the percentages of true positive and true negative scenarios that were classified with each response option and visualization.

Table 1. Percentage of positive and negative cases classified with each response option.

	Scenario type	**certainly yes**	**probably yes**	**maybe yes**	**maybe no**	**probably no**	**certainly no**
2D	Positive	25,1	39,9	7,4	6,9	11,3	9,4
	Negative	2,8	9,7	4,5	5,2	24,4	53,4
3D	Positive	22,0	37,0	9,5	8,0	19,5	4,0
	Negative	5,6	10,0	4,5	7,3	20,6	52,0

The percentages of true positive and true negative scenarios provide the basis for calculating the hit and false alarm pairs used for creating the ROC curves. Table 2 shows the results of cumulating the percentages beginning with the response option ‘certainly yes’ that are added to the values of the other response options beginning with ‘probably yes’ and so forth.

Table 2. Cumulated hit and false alarm rates over the response options beginning with ‘certainly yes’.

	Scenario type	**certainly yes**	**probably yes**	**maybe yes**	**maybe no**	**probably no**	**certainly no**
2D	Hit rate	25,1	65,0	72,4	79,3	90,6	100
	False alarm rate	2,8	12,5	17,0	22,2	46,6	100
3D	Hit rate	22,0	59,0	68,5	76,5	96,0	100
	False alarm rate	5,6	15,6	20,1	27,4	48,0	100

For illustrating the performances, the cumulated hit and false alarm rates for both the 2D and the 3D visualization are plotted into a coordinate system with the ordinate showing the hit rate and the abscissa the false alarm rate. Connecting all points including the zero scale marks result in the ROC curves shown in Figure 1a. For transforming the ROC curves into zROC graphs, first the hit and false alarm rates of Table 2 are transformed into standardized z-values by dividing them by 100 and consulting the respective z-values. Table 3 shows the results of this transformation. Please note that for the response option, for which the cumulated hit and false alarm rates amount to 100% per cent, no z-values can be reported, because z-values of the standard normal distributions range from - ∞ to + ∞.

Table 3. z-values of the cumulated hit and false alarm rates from Table 2.

	Scenario type	**certainly yes**	**probably yes**	**maybe yes**	**maybe no**	**probably no**	**certainly no**
2D	Hit rate	-0,67	0,39	0,59	0,82	1,32	---
	False alarm rate	-1,91	-1,15	-0,95	-0,77	-0,09	---
3D	Hit rate	-0,77	0,23	0,48	0,72	1,75	---
	False alarm rate	-1,59	-1,01	-0,84	-0,60	-0,05	---

For these values linear regressions are calculated. In the case of the 2D visualization, a slope of 1.1 and an intercept with the axis of ordinates of 1.56 within the z-score-based coordinate results. For the 3D visualization these values amount to 1.61 and 1.8 respectively. Both the calculation of the z-values as well as of the linear equations can be completed using commercial spreadsheet programs. Afterwards, the standard normal values of the hit and false alarm rates as well as the results of the linear equations are plotted into a coordinate system similar to the one used for displaying the ROC curves but with z-standardized axis. The zROC graphs are shown in Figure 1b, using a z-value-range from -2.5 (1%) to 2.5 (99%).

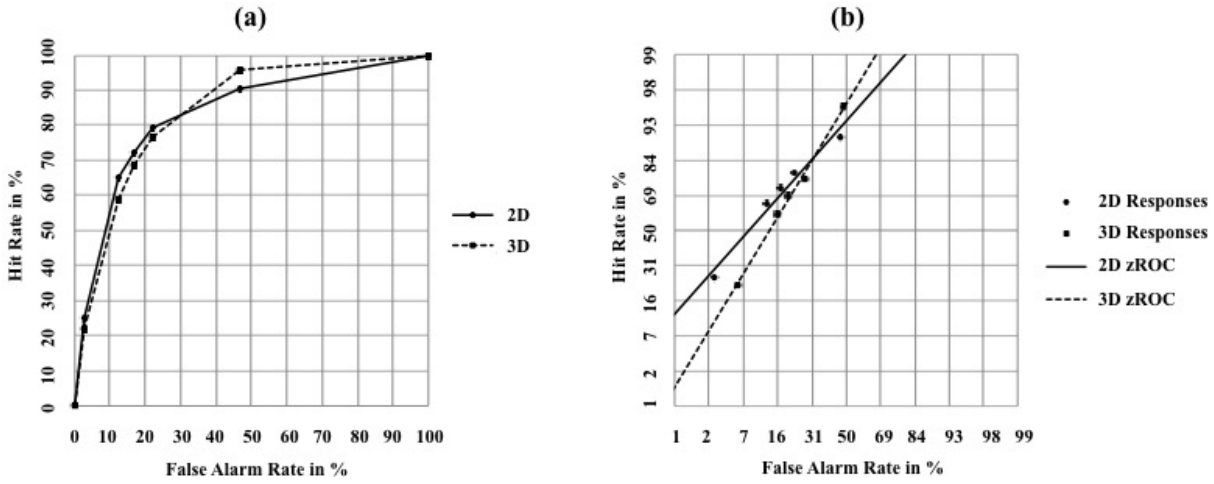

Figure 1. ROC curves (a) and zROC graphs (b) based on the hit and false alarm rates of the air traffic controllers with the 2D and the 3D visualization.

Which human-machine-interface leads to the best overall discrimination performance?

In order to evaluate the resulting discrimination performance with 2D and 3D, the areas under the both ROC curves shown in Figure 1a are compared. For the calculation of the AUC values, we refer to Green & Swets [1], because manually determining them is somewhat complex, and the description of the mathematical foundations required doing so would go beyond the scope of this article. We rather recommend using one of the various commercial statistics programs that offer the possibility to calculate AUC values, e.g. SPSS. Our results indicate that the use of the 2D visualization results in AUC value of 0.834 while the 3D visualizations leads to a result of 0.815. Hence, the likelihood for correctly classifying a randomly chosen case out of the 32 scenarios as conflict or separation is 83.4% when presented with 2D, and 81.5% in case 3D is used. Because the AUC values only refer to the size of the area under the ROC curves and neglect their shapes, this advantage of 2D over 3D is independent from the underlying judgment certainty, and reflects the average performance that is to be expected, no matter which criterion the air traffic controller decides to apply. This constitutes a major advantage over other methods for comparing the performance between human-machine-interfaces used for making binary choices, because the factors that impact on the deciders' judgment certainty and his or her decision about which criterion to apply in order to deal with uncertainties can be disregarded.

How does the response behaviour impact on performance?

In some cases the response criterion cannot be disregarded, but rather is of major importance. In air traffic control, for instance, the response behaviour is central, because safety is to be prioritized higher than efficiency, and the consequences of overlooking a conflict are worse than causing a false alarm. In this case, performance shall be measured by the amount of false alarms that result when the decider tends to favour the classification of uncertain cases as positives rather than negatives. Hence, the criterion by which the performance of the decider with different interfaces is compared matters. Transforming the ROC curves into zROCs allows the evaluator to choose the criterion by which the human-machine-interfaces shall be compared. In our example, either hit rates reported in studies from other researchers or the hit rates that resulted with the visualizations we evaluated constitute suitable reference values. The former allows for an invaluable comparison with other systems, while the latter offers a comparison between the traditional 2D top-view visualization currently used at air traffic controller workstations and the novel 3D visualization. Using the linear equation that describes the performance of the air traffic controllers with the 2D visualization, a false alarm rate of 57% is predicted for a criterion that leads them to classify 96.0% of the actual conflicts as such. This predicted value now could be compared with the false alarm rate of 48.0% that resulted with the 3D

visualization for the hit rate of 96.0%. The result indicates that by using the 3D visualization, a 9% lower false alarm rate can be expected compared with using the traditional 2D top-view when an equal conflict detection performance as with 3D shall be guaranteed. Please note that, because the linear equations are based on the z-transformed, values z-values have to be used for the calculations and that the result has to be converted into percentile ranks for its interpretation.

Interestingly, the result of comparing the false alarm rates between 2D and 3D shows the very revers result of the AUC comparison. While in the former comparison 3D turns out to be the advantageous visualization, the latter demonstrates 2D to be superior. The reason for this, as can be seen in Figure 1b, is the different slopes of the zROC. Therefore, the result depends on the response criterion of the decider and, in our example, the more liberal the criterion, the higher the advantages of 3D and vice versa. Hence, in other applications than air traffic control where it might be preferable to minimize the false alarm rate rather than maximizing the hit rate, e.g. because the costs of a false alarm outbalance those of missing a positive case, the application of conservative response behaviour is conveyed, and the same results would indicate 2D to be the preferred visualization. The reason for different zROC slopes lays in the variation of the deciders' responses when rating positive and negative cases. Figure 2 shows two examples of probability distributions that could result from rating scenarios on a six point Likert-scale. The graph to the right indicates the probability distribution that results from rating the positive scenarios, the left from rating the negative scenarios. In Figure 2a, an example is given in which the variation from the average value of the judgments is equal for both positive and negative cases. This leads to a unit slope of the zROC, because the growth of the probability for identifying a positive case as such when allowing more and more uncertainty (moving the criterion from the right hand side of the graph to the left hand side) increases in the same manner as the probability for a false alarm. The example depicted in Figure 2b shows two distributions with the same average values as the example in Figure 2a. Therefore, the discrimination performances of both examples are equal. In the example shown in Figure 2b, however, the standard deviation of the responses to the negative cases is larger than the deviation of the responses to the positive cases. Consequently, when applying a more conservative criterion, the probability for a false alarm initially is higher compared with the example of Figure 2a, but increases less when moving towards more liberal responses. When plotting both examples into a z-coordinate system, the example in Figure 2b therefore will result in a steeper zROC slope as the example shown in Figure 2a.

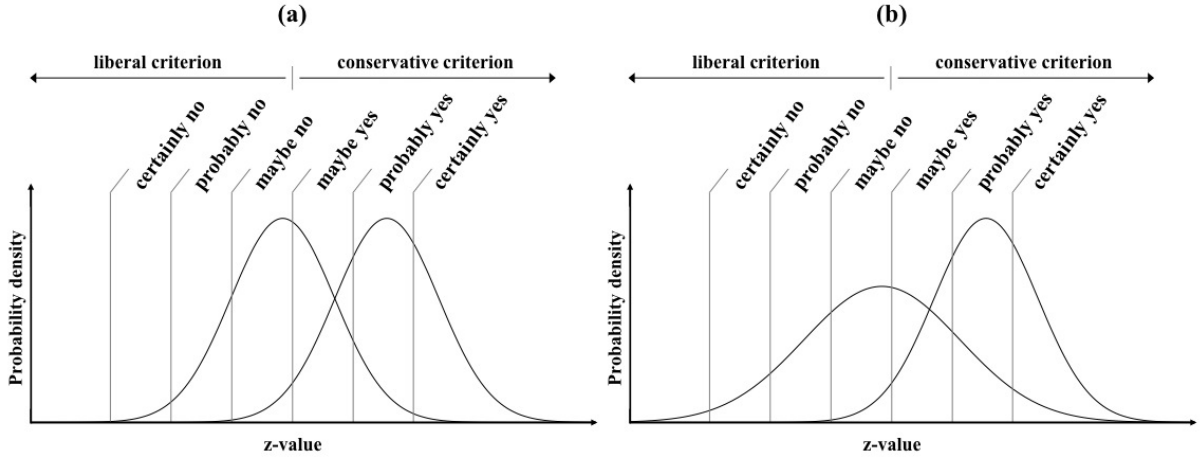

Figure 2. Exemplary distributions that result when the variances of the rating positive and negative cases are equal (a) or different (b).

In our air traffic control example, the positive and the negative scenarios were almost equal. While all factors such as horizontal and vertical aircraft speeds, directions, and approach angles were the same for both positive and negative cases, within the latter horizontal and vertical separations were created by separating their trajectories in the accordant direction. Hence, distinguishing a horizontal separation from a conflict only required the perception of the horizontal aircraft trajectories, whereas a vertical

separation could be discriminated from a conflict by processing the vertical aircraft trajectories alone. Because of the characteristics of the visualizations, the air traffic controllers were more certain when judging vertical separations with 3D, but less certain when horizontal separations where displayed. That is, their expertise with 2D visualizations vanishes in case of vertical separation.

Which response behaviour offers the best trade-off between hit and false alarm rate?

For some applications it is important to know the response criterion that offers the best trade-off between hit and false alarm rate. This might be the case when the binary choice is one of many in a process, and therefore optimizing the criterion does not impede the overall efficacy as could be the case in airport security scans. The best criterion can be determined by selecting the highest value that results from calculating the Youden-index [1], which is calculated by adding the sensitivity (hit rate) to the specificity (1 - false alarm rate) and subtracting one. In our example with the air traffic controllers, the best trade-off between hit and false alarms for both visualizations results, if all cases that fall in the response options from ‘certainly yes’ till ‘maybe no’ would be treated as conflicts and all cases that are classified with ‘probably no’ and ‘certainly no’ as separations.

DISCUSSIONS AND CONCLUSIONS

A common concern about using rating scales with more than two response options is that either before or after gathering the data, the evaluator has to define a criterion to decide which cases are positive and which are negative. This is of special concern when this criterion has to be chosen arbitrarily. The above-described procedures illustrate that by reporting the sizes of the areas under the receiver operating characteristic curves, no such decision is required for evaluating the discrimination performance. Though the area under the curve can be calculated on the basis of binary responses, using an appropriate Likert-scale offers several advantages for the evaluation of human-machine-interfaces. As stated above, the area under the curve offers measure of performance that is independent from judgment certainty, and allows for an objective comparison of the discrimination performance without the results being influenced by the deciders’ applying different response criteria with the human-machine interfaces as a reaction on differences regarding expertise or characteristics of the task or situation. Moreover, using a Likert-scale allows the evaluator to assess the performance for any criterion or choosing one by which the performances are compared. Also the determination of the most efficient response criterion is possible and can be used for training deciders in order to achieve the best trade-off between positive and false positive decisions. Above all, using a Likert-scale facilitates the deciders in rating the cases, because they are not forced to give a yes or no answer though they are uncertain.

REFERENCES

[1] Green, D.M., Swets J.A. (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley.
[2] Baier, A. (2013). *Stereoskopische 3D Anzeigen für die Flugsicherung*. Regensburg, Germany: Universität Regensburg.