Accounting for the Speed-accuracy Trade-off in Quantifying Human-in-the-loop Error Probabilities

Albert Ahumada[1] Bettina L. Beard[2]
NASA Ames Research Center, Moffett Field, CA, 94035

Cynthia Null[3]
NASA Langley Research Center, Hampton, VA, 23681

Human-in-the loop (HITL) simulations cannot collect enough data from human operators to validate estimates of error probabilities for task components. Error rates for tasks have been estimated by using laboratory data for error rates depending on variables such as the cognitive complexity of the task. The limited channel capacity of human operators compels error rates to be strongly related to the time available for the task, the speed-accuracy trade-off. HITL simulations can provide valuable data on the time available for the operator's tasks. We propose that the response times be used in conjunction with measured speed-accuracy curves to estimate the operator error rates contributing to mission failure. Such analyses should be especially important in the estimation of error rates in off-nominal situations.

Nomenclature

a, b = Beta distribution parameters
A = signal area
d' = signal-to-noise ratio
E = signal energy
f_0 = error probability prior distribution probability density
f_N = error probability posterior distribution probability density after N trials
F = cumulative distribution function
F_0 = error probability posterior cumulative distribution after N trials
F_N = error probability posterior cumulative distribution after N trials
Fn = standard normal cumulative distribution function
I = stimulus intensity level
N = number of trials
N_0 = prior distribution parameter
n_0 = noise power density
n_c = confidence level number = $1/(1 - p_c)$
$n_{e,0}$ = prior distribution error probability limit number = $1/p_{e,0}$
n_e = error probability number = $1/p_e$
$n_{c,0}$ = prior distribution error probability parameter number = $1/p_{e,0}$

[1] Research Psychologist, Human Systems Interface Division, NASA Ames Res. Ctr., Moffett Field, CA, 94035/Mail Stop 262-2, nonmember.
[2] Research Psychologist, Human Systems Interface Division, NASA Ames Res. Ctr., Moffett Field, CA, 94035/Mail Stop 262-4, nonmember.
[3] Technical Fellow for Human Factors, Human Systems Interface Division, NASA Ames Res. Ctr., Moffett Field, CA, 94035/Mail Stop 262-11, nonmember.

$p_{c,\,0}$ = prior probability that $p_e < p_{e,\,0}$
P_d = probability of a correct discrimination
P_I = probability of a correct identification
p_c = confidence level as a probability
p = discrete probability distribution function
p_e = error probability
$p_{e.\,0}$ = prior distribution error probability limit
s = sensory noise standard deviation
s_0 = sensory pattern filtered noise standard deviation
t = stimulus duration
X = number of trials resulting in errors

I. Introduction

One possible goal of HITL simulations is to provide data to evaluate the system reliability. Human operator reliability might be assessed by the frequency that errors are made in the simulation. Here we point out that even if no errors are made, the required number of opportunities for error may be impractically large if a high level of reliability must be insured. We propose using the strategy used in the perceptual sciences to ensure that stimuli can be discriminated with high reliability. Measurements are made with stimuli that lead to a measurable rate of confusion and then the low error-rate stimuli are derived from simple models. Most cognitive tasks involve a time-accuracy trade-off, so that measurable accuracy can be achieved by making the available time short enough. Part-task simulations to measure the time-accuracy trade-off can be made more efficient through the use of sequential designs efficiently designed to estimate model parameters. Standard HITL simulations are then seen, not as a source of error rates, but as a source of distributions of times available for cognitive operations, which can then be fed into time-accuracy trade-off models to obtain error-rate estimates.

II. The error rate measurement problem

In the performance of critical tasks, system designers would like to have statistical evidence that the probability of human error will be less than some value $p_e = 1/n_e$ with some level of confidence $p_c = 1 - 1/n_c$. For example, n_e might be 100 and n_c might be 10.

For values of $n_e > 10$, it is shown in the Appendix that even if the HITL simulations opportunities resulted in no errors, the number of trials needed N would be

$$N = n_e \ln n_c = 2.3\, n_e \log n_c . \qquad (1)$$

For n_e =100 and n_c = 10, we obtain N = 230.

Notice that if we want the error rate to be 10 times less, p_e = 1/1000, we need 10 times as many trials, but that if we want the confidence number to be 10 times more so that the confidence is 0.99 instead of 0.90, we only double the number of trials needed.

The Appendix also shows that if you have the prior belief that the probability is $p_{c,\,0}$ that the error rate is less than $p_{e,\,0}$, the number of error-free trials N needed to allow the belief that it is less than p_e with a probability of p_c, is

$$N = \log(1 - p_e)/\log(1 - p_e) - N_0 \approx 2.3\, n_e \log n_c - N_0 . \quad (2)$$

where

$$N_0 = \log(1 - p_{c,0}) / \log(1 - p_{e,0}) \approx 2.3\, n_{e,0} \log n_{c,0}, \quad (3)$$

and $n_{e,0}$ and $n_{c,0}$ have the expected meanings, $n_{e,0} = 1/p_e$ and $n_{c,0} = 1/(1 - p_c)$. The two prior assumption probabilities can be regarded as providing a number of trials that do not need to be run.

III. Modeling Errors

We propose that rather than measure human error rates directly, that they be estimated using models. Sensory systems can be regarded as a cascade of processes, where each level contributes its own noise.[1] In the visual system at low levels, the photon noise can be dominant, but at normal working levels the neural noise from the retinal ganglion cells is regarded as the noise that limits performance.[2-4] This is plausible since the information from 10^7 cones must carried to the brain on 10^6 nerve fibers using an inherently noisy pulse code. In the brain, resources can be devoted to reduce further degradation. A useful model for visual discrimination errors is that the visual signals are approximately linearly filtered in the retina and that spatio-temporal white noise is added by the ganglion cells.[2] This signal is relayed to the cortex where an ideal statistical analysis is performed for the task required.[5] The result of this analysis is a single number which has variability as a result of the noise and inherently results in sensory errors. When there are no uncertainties and the white noise is Gaussian, the ideal analysis is based on the cross correlation of the signal pattern in the domain where the noise is added with noisy signal. This results in the detectability of a signal being determined by the signal energy after the retinal filtering.[6] Signal detection theory shows that the ideal observer performance is characterized by

$$d' = \sqrt{(E / n_0)}, \quad (4)$$

where E is the signal energy and n_0 is the noise spectral density. In vision the signal is usually regarded as a contrast signal decomposable into the product of a peak RMS contrast C, an energy equivalent area A, and energy equivalent duration t. The detectability d' relates to these variables as

$$d' = C \sqrt{(A\, t)} / s_0, \quad (5)$$

where s_0 is the standard deviation of the noise resulting from the cross correlation of the noise with the signal.[5]

1. Sensory errors, sensory thresholds, and JND's

When psychologists measure sensory thresholds they assume that the trial-to-trial variations in task performance represent the effect of sensory noise. Suppose the goal is to measure the threshold for the difference in stimulus intensity at intensity level I_0. On each trial two stimuli are presented, I_0 and I_1. The probability of a correct detection can be modeled as

$$P_d = Fn\,(I_1 - I_0)/(s\sqrt{2}), \quad (6)$$

where Fn is the standard normal cumulative distribution function and s is the standard deviation of the sensory noise on a single presentation, I_0 or I_1. Measurements of discrimination are usually reported as the size of a just-noticeable-difference (JND) in the stimulus domain. This may be defined as the intensity difference $I_1 - I_0$ that leads to 75% correct responses. Equivalently, the experiment can be regarded as measuring σ at I_0, since

$$s = (I_1 - I_0)\, Fn^{-1}(0.75) / \sqrt{2} = 0.477\,(I_1 - I_0). \quad (7)$$

To ensure that sensory noise does not limit the discriminability of colors, for example, colors are often chosen to be at least 10 JNDs apart. The probability that Gaussian noise would be large enough to cross a boundary 5 standard deviations away is about 3×10^{-6}. Notice that if we wanted to perform an experiment to ensure with 95% confidence that the error rate was less than 10^{-5} we would need to make over 45,000 error-free observations.
Table1 Shows distances in JNDs using the Lab color metric for colors being considered for use in control room displays.[7]

JNDs between colors using the Lab color metric.

RGB Input				Normal Color Vision								
Colors	**R**	**G**	**B**	*Red*	*Orange*	*Yellow*	*Green*	*Cyan*	*Purple*	*Black*	*White*	*Gray*
Red	255	0	0	--								
Orange	255	160	80	42	--							
Yellow	255	255	0	79	44	--						
Green	0	192	131	112	75	72	--					
Cyan	0	185	255	128	96	109	51	--				
Purple	175	0	255	115	104	138	51	69	--			
Black	0	0	0	101	92	116	79	85	91	--		
White	255	255	255	94	58	69	53	51	81	100	--	
Gray	130	130	170	94	68	92	51	36	56	61	44	--
Dark Gray	75	75	75	87	67	92	53	59	74	33	67	31

RGB Input				Protanope								
Colors	**R**	**G**	**B**	*Red*	*Orange*	*Yellow*	*Green*	*Cyan*	*Purple*	*Black*	*White*	*Gray*
Red	255	0	0	--								
Orange	255	160	80	24	--							
Yellow	255	255	0	54	35	--						
Green	0	192	131	31	21	54	--					
Cyan	0	185	255	73	67	96	46	--				
Purple	175	0	255	95	94	127	46	34	--			
Black	0	0	0	60	80	113	72	83	82	--		
White	255	255	255	65	47	66	34	37	71	100	--	
Gray	130	130	170	54	53	87	32	23	42	62	43	--
Dark Gray	75	75	75	39	53	88	40	53	63	33	67	30

RGB Input				Deuteranope								
Colors	**R**	**G**	**B**	*Red*	*Orange*	*Yellow*	*Green*	*Cyan*	*Purple*	*Black*	*White*	*Gray*
Red	255	0	0	--								
Orange	255	160	80	21	--							
Yellow	255	255	0	34	27	--						
Green	0	192	131	55	42	68	--					
Cyan	0	185	255	105	90	114	50	--				
Purple	175	0	255	120	108	132	50	20	--			
Black	0	0	0	88	92	118	66	82	82	--		
White	255	255	255	74	54	70	35	51	71	100	--	
Gray	130	130	170	80	68	94	26	27	40	61	45	--
Dark Gray	75	75	75	69	66	93	33	56	63	33	67	31

The goal was to find colors that were at least 30 JNDs apart for the three color types that capture the sensitivity of the common anomolous types, deuteranomaly and protananomaly. This distance seems extreme, but the calibration experiments used stimuli with a diameter of 2 degrees of visual angle and the colored areas in the display are much smaller. If the linear dimension is reduced to only 0.2 degrees of visual angle, the JND criterion would then only be 3 JND's. One problem with trying to base the estimated sensitivity on area rather than actual ganglion cell density is that the ganglion cell density associated with the blue cones is nearly zero in the center of the fovea, leading to small field tritanopia, where the white and the yellow are not distinguishable.

Another good example of modeling human error by sensory noise is the pattern vision model, which assumes that the visual stimulus is filtered by the optics, converted to contrast in the eye and then has spatio-temporal white noise added to it. The detection or discrimination processing is then assumed to be ideal. Once the filter properties have been measured, any experiment in pattern discrimination allows estimation of the noise level. In this model, if a stimulus is just at threshold, it can be moved a predictable distance in JNDs from threshold by increasing the area or the duration. The changing the duration changes the signal-to-noise according to the formula,

$$d' = d'_0 \sqrt{(t / t_0)}, \quad (8)$$

where d' is the new signal-to-noise ratio, d'_0 is the original one, t is the new duration and t_0 is the original one. If error rates are measured at short durations, this model of the human observer allows the prediction of sensory noise based error rates for longer duration stimuli, rates that are too low to be actually measured.

When observers are allowed to control their own speed, this noise model allows prediction of the speed-accuracy trade-off that results. We express the signal-to-noise ratio as

$$d' = d'_1 \sqrt{t}, \quad (9)$$

where d'_1 is the signal-to-noise ratio for a signal duration of one second. Fig. 1 shows the results from an experiment in letter identification.[8]

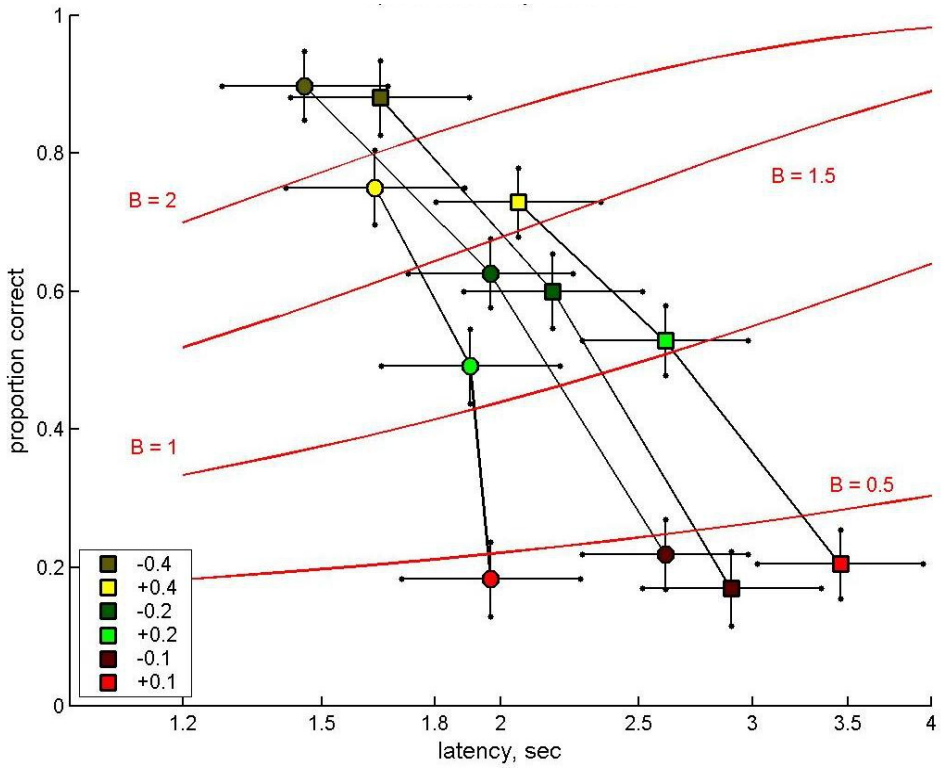

Figure 1. Letter identification accuracy vs. latency (speed-accuracy trade-off). *Square symbols indicate conditions run in first group of 5 blocks. Circles indicate second group of 5 blocks. Colors indicate the contrast in percent. Error bars are 95% confidence intervals based on the observer x treatment interaction. Red lines show constant performance curves for our speed-accuracy trade-off model. The parameter* B *is the parameter* d'_1 *in the text, the signal-to-noise ratio at a latency of one second.*

Under the simplifying assumption that the differences among the 12 letters are orthogonal stimuli, a signal-to-noise ratio of d' should lead to a probability of a correct identification P_I approximately given by

$$P_I = Fn\,(0.87\,d' - 1.38), \quad (10)$$

where *Fn* is the cumulative standard normal distribution and the constants are from Elliott.[9] There is a striking difference between the latencies (speed) for the positive contrast conditions (light colored symbols) depending on whether they were first (squares) or second (circles). The model suggests that although the student subjects seemed to be more in a hurry to leave in the second session, they were actually performing slightly better despite the proportion of correct responses being slightly lower.

2. Cognitive errors

Speed-accuracy trade-off experiments in cognitive decision making have used this same basic model, assuming that the subject is accumulating noisy information over time.[10]

Conclusion

Human-in-the loop (HITL) simulations cannot collect enough data to validate estimates of error probabilities for task components by human operators. Error rates for tasks have been estimated by using laboratory data for error rates depending on variables such as the cognitive complexity of the task. The limited channel capacity of human operators compels error rates to be strongly related to the time available for the task, the speed-accuracy trade-off. HITL simulations can provide valuable data on the time available for the operator's tasks. We propose that the times be used in conjunction with measured or theoretical speed-accuracy curves to estimate the operator error rates contributing to mission failure. Such analyses should be especially important in the estimation of error rates in off-nominal situations.

Appendix

If an experiment provides N independent and equally probable opportunities for an error, the probability that there will be X errors in N trials when the probability on a single trial is p_e, is given by the binomial distribution

$$p(X; N, p_e) = (N!/(X!\,(N-X)!)\, p_e^{\,X} (1 - p_e)^{N-X} \quad . \quad \text{(A1)}$$

The upper limit of the one sided confidence interval for p_e with confidence level p_c is given by the smallest value of p_e that can be rejected by the experiment at a significance level of $1\text{-}p_c$. The cumulative distribution of X, $F(X; N, p_e)$ is the probability that no more than X errors will occur,

$$F(X; N, p_e) = \sum_{x=0}^{X} p(x; N, p_e) \quad . \quad \text{(A2)}$$

The confidence limit is thus given by the value p_e such that

$$F(X; N, p_e) = 1 - p_c \, . \quad \text{(A3)}$$

Since the binomial distribution functions are readily available, it is straightforward to solve this equation numerically.

In the case that no errors are observed and $X = 0$,

$$F(0; N, p_e) = p(0; N, p_e) = (1 - p_e)^{N}, \quad \text{(A4)}$$

and the confidence limit occurs when

$$1 - p_c = (1 - p_e)^{N} \, . \quad \text{(A5)}$$

Solving for p_e, we obtain

$$p_e = 1 - (1 - p_c)^{1/N} \qquad \text{(A6)}$$

If we have desired values for p_c and p_e, we can solve for N

$$N = \log(1 - p_c) / \log(1 - p_e). \qquad \text{(A7)}$$

For small p_e ,

$$\log(1 - p_e) = \ln(10) \ln(1 - p_e) \approx - \ln(10) , \qquad \text{(A8)}$$

so

$$N = - \ln(10) \log(1 - p_c) / p_e = 2.3 \, n_e \log(n_c), \qquad \text{(A9)}$$

where $n_e = 1/p_e$ and $n_c = 1/(1 - p_c)$.

If we want to be 99% confident (n = 100) that the error rate is less than 0.01 (n_e = 100), we find that the number of observations N must be 459 by the exact formula or 2.3 x 2 x 100 = 460, by the approximation. Note that this number is not very sensitive to the confidence level, because of the log, but gets large quickly if the error probability must be very small, i.e. if we want to know that the operator is very reliable.

With a Bayesian approach, the result is simplified if the priori distribution is chosen to be the Beta distribution. In general, if the prior distribution of p_e is Beta with parameters a and b, the posterior distribution after a binomial experiment with X errors in N trials will be Beta with parameters $a + X$ and $b + (N-X)$.

If p_e is known to be small, it is convenient to choose prior distribution parameters a = 1 and $b = N_0$, so that

$$f_0(p_e) = N_0 (1 - p_e)^{N_0 - 1} . \qquad \text{(A10)}$$

This distribution has a mean of $1/(N_0 +1)$ and a standard deviation of $1/((N_0 +1) \sqrt{(1+2/N_0)}$.

The cumulative distribution is

$$F_0(p_e) = 1 - (1 - p_e)^{N_0} \qquad \text{(A11)}$$

If our prior knowledge is that we think that the probability is $p_{c,\,0}$ that p_e is less than $p_{e,\,0}$, then we need to choose N_0 so that

$$p_{c,\,0} = 1 - (1 - p_{e,\,0})^{N_0} , \qquad \text{(A12)}$$

which is

$$N_0 = \log(1 - p_{c,\,0}) / \log(1 - p_{e,\,0}) \approx 2.3 \, n_{e,\,0} \log(n_{c,\,0}). \qquad \text{(A13)}$$

When the experimental result of no errors in N trials is used to update this prior, the resulting posterior distribution has the same form with N_0 replaced by $N_0 + N$,

$$f_N(p_e) = (N_0 + N) (1 - p_e)^{N_0 + N} . \qquad \text{(A14)}$$

The cumulative posterior distribution is given by

$$F_N(p_e) = 1 - (1 - p_e)^{N_0 + N} . \qquad \text{(A15)}$$

The value of N that ensures that the probability is $p_{c,\,N}$ that p_e is less than $p_{e,\,N}$, satisfies

$$N + N_0 = \log(1 - p_{c,\,N}) / \log(1 - p_{e,N}) \approx 2.3 \, n_{e,\,N} \log(n_{c,\,N}). \qquad \text{(A16)}$$

The quality of the approximation is shown computing the proportional error

$$e = (\ln(1 - 1/N) - N)/N. \quad \text{(A17)}$$

For $N = 11$, $e = 0.0087$; and for $N = 100$, $e = 0.00010$.

Acknowledgments

This work was supported by the NASA Engineering and Safety Center.

References

[1]Watson, A. B. "Transfer of contrast sensitivity in linear visual networks," *Visual Neuroscience,* Vol. 8, pp. 65-76, 1992.

[2]Barton, P. G. J. "Spatio-temporal Model for the Contrast Sensitivity of the Human Eye," *Dynamic Properties of Vision*, Vol. VII, edited by F. Engel, H. de Ritter, F. Blommaert, Institute for Perception Research, Eindhoven, The Netherlands, 1994, pp. 33-46.

[3]Warland, D. K., Reinagel, P., Meister, M. "Decoding visual information from a population of retinal ganglion cells." *Journal of Neurophysiology,* Vol. 78, No. 5, pp. 2336-2350, 1997.

[4]Watson, A. B., "A formula for human retinal ganglion cell receptive field density as a function of visual field location," *Journal of Vision,* Vol. 14, No. 7, Paper 15, 2014.

[5]Watson, A. B., Barlow, H. B., Robson , J. G.(1983) "What does the eye see best?" *Nature* , Vol. 302, No. 5907, pp. 419-422, 1983.

[6]Ahumada, A. J., Watson, A. B. (2013). "Visible contrast energy metrics for detection and discrimination," *Human Vision and Electronic Imaging XVIII*, edited by B. E. Rogowitz and T. N. Pappas., SPIE Proceedings 8651, 2013, paper 13.

7Pauli, H. "Proposed extension of the CIE recommendation on 'Uniform color spaces, color difference equations, and metric color terms,' " *Journal of the Optical Society of America*, Vol. 66, No. 8, pp. 866-867, 1976.

[8]Ahumada, A. J., Scharff, L. F. V., "Letter Identification: Contrast Polarity and Speed-Accuracy Trade-Off Strategies," *Abstracts of the Psychonomic Society 44th Annual Meeting*, Vol. 8, 2003, p. 67.

[9]Elliott, P. B. "Tables of d'," *Signal Detection and Recognition by Human Observers*, edited by J. A. Swets,., Wiley, New York, 1964, pp. 651-684.

[10]Usher, M., McClelland, J. L. "The time course of perceptual choice: The leaky, competing accumulator model," *Psychological Reviews*, Vol. 108, No. 3, 2002, pp. 550-592.