USING SENSITIVITY AND BIAS IN SIGNAL DETECTION THEORY TO PREDICT PROPORTION CORRECTNESS: SIMULATION AND CASE STUDY ON ADAS MENTAL MODEL EVALUATION

Chunxi Huang[1], Dengbo He[2,1,3]
1. Robotics and Autonomous Systems Thrust, The Hong Kong University of Science and Technology (Guangzhou)
2. Intelligent Transportation Thrust, The Hong Kong University of Science and Technology (Guangzhou)
3. HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen

The sensitivity (*d'*) and bias (*c*) in signal detection theory can reflect respondents' objective performance in understanding a system and their bias towards saying yes. Thus, *d'* and *c* can be used as alternatives of proportion correctness (PC) when evaluating drivers' mental models of advanced driving assistance system (ADAS). The adoption of *d'* and *c* as mental model metrics also allows cross-study comparisons as their values are independent of signal (i.e., ADAS function present) and noise (i.e., ADAS function absent) ratio. However, there is no closed-form solution of the relationships among *d'*, *c*, and PC. Hence, using numerical simulations, we extracted an empirical equation that quantifies how *d'* and *c* can estimate PC. The equation was then validated based on participants' responses from a survey study that targeted towards drivers' ADAS mental model. The results show that the empirical equation reached a satisfying performance ($R^2 > 0.8$) in estimating PC.

INTRODUCTION

The past few years have witnessed an increasing number of ADAS-related crashes (National Highway Traffic Safety Administration, 2022), most of which are due to or partially due to drivers' misusing or overusing of advanced driver assistance systems (ADAS). Previous studies found that drivers who have inappropriate mental models of ADAS (i.e., understanding of the functions and limitations of ADAS) were more likely to misuse or overuse the systems (Dickie & Boyle, 2009; Rossi et al., 2020). Hence, researchers tried to understand the factors influencing drivers' mental model of ADAS, which can provide insights on how to help drivers construct appropriate ADAS mental models.

To understand drivers' mental model of ADAS, most previous studies used survey-based method. Usually, in the survey, participants were presented with multiple statements about whether an ADAS function/limitation was present or not. Then, two types of responses were collected, i.e., binary answers (i.e., yes or no response indicating whether a function/limitation exists) or rating answers (i.e., Likert scales indicating how confident participants feel that the function/limitation exists). To evaluate the mental models, proportion correctness (PC) has been widely adopted in previous studies (for instance, a participant would get a PC of 45% if 45 out of 100 questions were answered correctly). DeGuzman and Donmez (2021) evaluated drivers' understanding of adaptive cruise control (ACC) and lane-keeping assistant (LKA) through an online survey consisted of 51 questions, in which participants were asked to judge whether the statement regarding the functions and limitations of ACC/LKA was true or false. Similarly, Larsson (2012) assessed drivers' understanding of the ADAS based on their "yes or no" answers to statements regarding ADAS functions and limitations. The performance of the mental models in the previous two studies was based on the PC out of all questions used in the surveys. When rating scale answers were collected to evaluate participants' mental model performance, a more complex calculation procedure was adopted to calculate the PC. For example, Beggiato and Krems (2013) assessed drivers' knowledge of ACC using a 35-question survey, in which respondents answered questions using a Likert scale ranging from 1 ("fully disagree") to 6 ("fully agree"). The absolute difference between actual responses and correct answers was used to derive a PC-based score to indicate participants' mental models of ACC.

However, PC may be a biased evaluation of the mental model, especially when the number of signal (i.e., ADAS function present) and noise (i.e., ADAS function absent) is unbalanced. For example, a respondent who trusts ADAS more (thus more likely to believe a ADAS function exists) may obtain higher PC-based score if most of the statements are about existing ADAS functions compared to that when most of the statements are about the non-existing functions. To resolve this issue, given that drivers' responses are binary (yes or no) or rating scales, and the signals are either present or absent, the signal detection theory (SDT) (Green & Swets, 1966) may be a better choice when measuring the mental model. In SDT, sensitivity (*d'*) can provide unbiased measure of drivers' understanding of the mental model objectively and response bias (*c*) can reflect drivers' subjective bias in believing whether the signals are present or not. Given that previous research mostly used PC-based score to measure the ADAS mental model, to facilitate comparisons across different studies and future meta-analysis, it is necessary to reveal the relationships among *d'*, *c*, and PC. However, to the best of knowledge, no research has clearly provided such relationships, given that there is no closed-form solution.

Therefore, this study aims to quantify the relationships among *d'*, *c*, and PC when binary answers or rating answers were collected in the mental-model-related surveys. Empirical equations were extracted through numerical simulations. To validate the extracted relationships, we further compared the

PC calculated from the empirical equation and the PC derived from a real survey study targeted towards evaluating drivers' ADAS mental model.

METHOD

Data Description

The data used in this study included two parts: (1) data acquired from simulation (Simulated Data); (2) data extracted from a survey study that explored drivers' understanding of ADAS technologies (Survey Data).

Simulated Data. We followed three steps to generate the Simulated Data:

(1) Generate a predefined perfect answer (PA) for *N* questions (i.e., a 1*N vector). The number of statements with only noise and the number of statements with signals were controlled by pre-set probability of noises, *p(n)*.

(2) Generate participants' responses (PR) to the *N* questions from *P* participants (i.e., a P*N matrix).

(3) Run the simulation *IterNo* (number of iterations) times.

Specifically, for questions with binary answers (i.e., binary questions), the predefined PA was either 0 (i.e., noise, or signal absent) and 1 (i.e., signal present). PR was also either 0 (i.e., "no, there is no signal") or 1 (i.e., "yes, there is a signal"), with the possibility of 50% for each. For questions with rating answers (i.e., rating questions), the PA was either 1 (i.e., "signal absent") or 6 (i.e., "signal present"). While participants' responses can be any integers from 1 (i.e., "I am very confident there is no signal") to 6 (i.e., "I am very confident that there is a signal"), with the possibility of 1/6 for each rating. Both PA and PR were generated using the '*sample'* function in R.

As the total number of trials (*N*), probability of noises (*p(n)*), and the number of participants (*P*) vary in different studies, to fully reveal the relationships among *d'*, *c* and PC, we simulated different combinations of these variables. Specifically, we set *N* = (100, 200, 300, 400, 500); *p(n)* = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9); *P* = (100, 200). It should be noted that, the exact number of noises in each iteration varied around *N*p(n)* due to the randomized sampling process. Therefore, we further defined the signal-to-noise ratio (SN-ratio) as the actual number of signals divided by the number of noises in each iteration. Most of SN-ratios distributed within the range of (0, 10) given the *p(n)* we set. In total, we simulated 90 conditions (5 N levels * 9 *p(n)* levels *2 P levels), and the number of iterations (*IterNo*) for each condition was 100. All the Simulated Data was generated using R 3.6.3 (R Core Development Team, 2019).

Survey Data. The data in this part was extracted from a survey study, in which we explored drivers' ADAS mental models using 49 statements about the functions and limitations of ADAS. The statements were generated based on a review of previous relevant studies (Beggiato & Krems, 2013; Beggiato, Pereira, Petzoldt, & Krems, 2015; DeGuzman & Donmez, 2021a, 2021b; McDonald, Carney, & McGehee, 2018) and user manuals from vehicle manufacturers. Participants were asked to rate their level of agreement to each statement, ranging from 1 ("*strongly disagree*") to 6 ("*strongly agree*"). Further, participants were informed that ratings below or equal to 3 indicated an intention to disagree while ratings above or equal to 4 indicated an intention to agree. Thus, we could also generate binary answers from this survey (i.e., if ratings were ≤ 3, then we treat the answer as "no, signal is absent"; while if ratings were ≥4, we treat the answer as "yes, signal is present"). Among all 49 statements, 34 of them can be treated as signals in SDT as they stated ADAS functions or limitations that indeed exist; while 15 of them can be treated as noises as they stated functions or limitations that do not exist. In total, 287 valid responses from the survey were used for analysis as Survey Data in the current study.

Variable Extraction

Dependent Variable. The PC was selected as the dependent variable for both binary questions and rating questions. More specifically, the PC of binary answers (PC-binary) was calculated as follows:

$$PC - binary\ = (\sum_{i\,=\,1}^{N} I(\mathrm{PR}_i, \mathrm{PA}_i)\,) * \frac{1}{N} \qquad (1)$$

where, $I\ (\mathrm{PR}_i, \mathrm{PA}_i) = \begin{cases} 1 & when\ \mathrm{PR}_i\ = \mathrm{PA}_i \\ 0 & when\ \mathrm{PR}_i\ \neq \mathrm{PA}_i \end{cases}$; PR_i was the actual response to the i^{th} question from a participant; PA_i was the perfect answer to the i^{th} question.

At the same time, the PC of rating answers (PC-rating) was calculated as follows:

$$PC - rating\ = (\sum_{i\,=\,1}^{N} \frac{5\ -\ |PR_i\ -\ PA_i|}{5}) * \frac{1}{N} \qquad (2)$$

where, PR_i was the actual response to the i^{th} question from the participant; PA_i was the perfect answer to the i^{th} question (i.e., either 1 or 6). Thus, being an integer within [0, 5], $|PR_i\ -\ PA_i|$ is the distance between the participant's response and the perfect answer for the i^{th} question.

Independent variables. The sensitivity (*d'*) and response bias (*c*) in SDT were selected as the independent variables. As both PA and PR are binary for binary questions, the traditional SDT (Stanislaw & Todorov, 1999) was adopted. We defined the hit as "signal present and answer yes", miss as "signal present but answer no", false alarm as "signal absent but answer yes", and correct rejection as "signal absent and answer no". For the rating questions, we adopted the fuzzy SDT to handle the non-binary responses. The fuzzy SDT enabled the analysis of fuzzy signals based on signal detection theory without arbitrarily binarizing the original responses (Parasuraman, Masalonis, & Hancock, 2000). In fuzzy SDT, the four states of the world (i.e., hit, miss, false alarm, and correct rejection) were calculated as follows:

$$Hit = \min(r, s) \qquad (3)$$

$$Miss = \max(s - r, 0) \qquad (4)$$

$$False\ Alarm = \max(r - s, 0) \qquad (5)$$

$$Correct\ Rejection = \min(1 - s, 1 - r) \qquad (6)$$

where, $r = \frac{5\ -\ |PR_i\ -\ PA_i|}{5}$ (i.e., the confidence of saying "yes"); s represents the probability of a signal (which was 0 when the signal was absent and 1 when the signal was present). The calculation of *d'* and *c* for both binary questions and rating questions was performed using the *psycho* package in R 3.6.3.

Extraction of Empirical Equation

Overall, the analyses in this study included two parts. First, by using the Simulated Data, we fitted linear regression models to explore the empirical relationships among *d'*, *c* and PC. Specifically, for each single iteration in one condition, we could calculate P sets of *d'*, *c* and PC, as we have P participants. Then, we were able to construct a linear model using these P sets of *d'* and *c* and PC. In the model, as shown in Eq. (7), *d'* and *c* as well as their two-way interactions were independent variables, and the corresponding PC was the dependent variable:

$$PC = \beta_0 + \beta_1 * d' + \beta_2 * c + \beta_3 * (d'|c) \qquad (7)$$

Where, $(d'|c)$ stands for the interaction between *d'* and *c*. For each condition, this procedure was repeated *IterNo* times. Further, we considered two types of questions (i.e., binary questions or rating questions). Thus, 18,000 models were fitted (90 conditions * 2 types of questions * 100 iterations).

In total, 18,000 sets of coefficients ((i.e., β_0, β_1, β_2, and β_3) in Eq (7) were generated. Then, to extract an empirical equation, in Figure 1 to 5, we visualized the distribution of these coefficients. Then, the extracted empirical equation was applied to the Survey Data. Specifically, the relationship between PC calculated from the empirical equation (predicted PC) and the PC derived from participants' responses (extracted PC) was fitted in linear regression models and the R-squared (R^2) was used as the evaluation metric. The model fitting process, data pre-processing process, and results visualization in this study were performed using R 3.6.3.

RESULTS

Relationships among *d'*, *c* and PC in Simulated Data

Summary of fitted models. Table 1 summarizes the descriptive statistics (i.e., mean, min, and max) of the coefficient estimates in all fitted models. It should be noted that the R-squared for all fitted models was over 0.999, indicating a close-to-perfect performance of using *d'* and *c* to predict PC.

Table 1. Descriptive statistics for all the estimates in all fitted models

Question Type	*Coefficient*	*Mean (SD)*	*Min*	*Max*
Binary questions	β_0	.50 (<.0001)	0.4998	0.5002
	β_1	.20 (.002)	0.1998	0.2027
	β_2	.0003 (.21)	-0.3754	0.3753
	β_3	<.0001 (.002)	-0.0169	0.0152
Rating questions	β_0	.50 (<.0001)	0.4993	0.5006
	β_1	.20 (.0004)	0.1982	0.2019
	β_2	<.0001 (.21)	-0.3757	0.3758
	β_3	<.0001 (.004)	-0.0307	0.0271

Note: SD stands for standard deviation.

Estimates of β_0. Figure 1 visualizes the relationships among N, P, SN-ratio, and estimates of β_0 for both binary questions and rating questions. It can be observed for both binary and rating questions, β_0 fluctuated around 0.5 (see Figure 1 and Table 1) regardless of the values of P and SN-ratio. At the same time, the intercept term (i.e., β_0) was always significant in all fitted models. Further, with the increase of N, β_0 tended to converge to 0.5. Therefore, we set β_0 = 0.5 in our empirical equation.

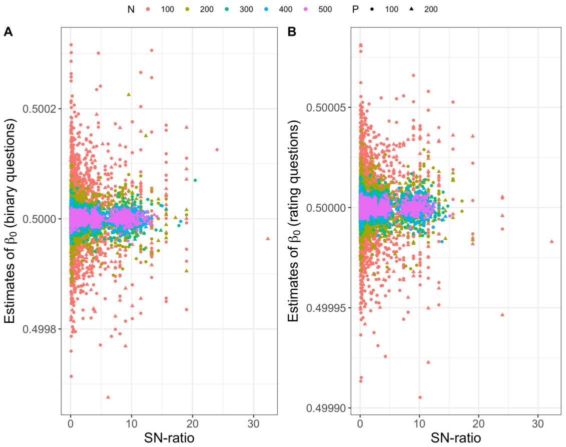

Figure 1. Relationships among N, P, SN-ratio, and estimates of β_0 for two types of questions: (A) binary questions; (B) rating questions.

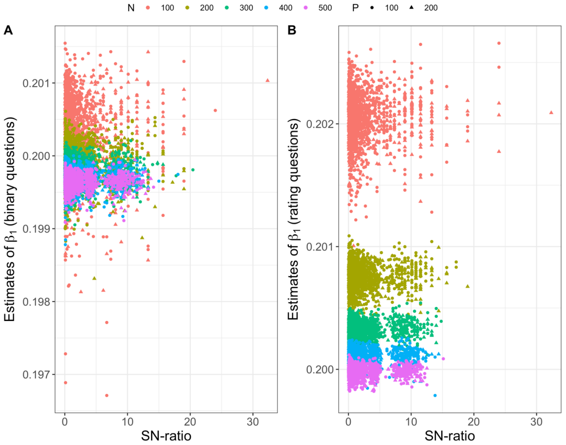

Figure 2. Relationships among N, P, SN-ratio, and estimates of β_1 for two types of questions: (A) binary questions; (B) rating questions.

Estimates of β_1. As shown in Figure 2 and Table 1, the β_1 fluctuated around 0.2 for different P and SN-ratio and the estimates of β_1 got closer to 0.2 when the N increased. At the same time, β_1 was always significant in all fitted models. Thus, we set β_1 as 0.2 in our empirical equation.

Estimates of β_2. From Figure 3, it can be observed that neither N nor P affects β_2. However, it is interesting to notice that the β_2 was a function of SN-ratio for both binary and rating questions and β_2 was always significant in all fitted models. Thus, to further quantify the relationship between β_2 and SN-ratio, inspired by the way of calculating c in SDT (i.e., transferring probabilities to z-score) and the curve shape in Figure 3, we fitted a function (Eq. (8)) using the nonlinear least squares algorithm in R (i.e., '*nls*' function) to represent the relationship between β_2 and SN-ratio. The fitted function was visualized in a curve in Figure 4 along with the simulated data points.

$$\beta_2 = \frac{e^{-0.73ln(SN-ratio)}}{1+e^{-0.73ln(SN-ratio)}} - 0.5 \qquad (8)$$

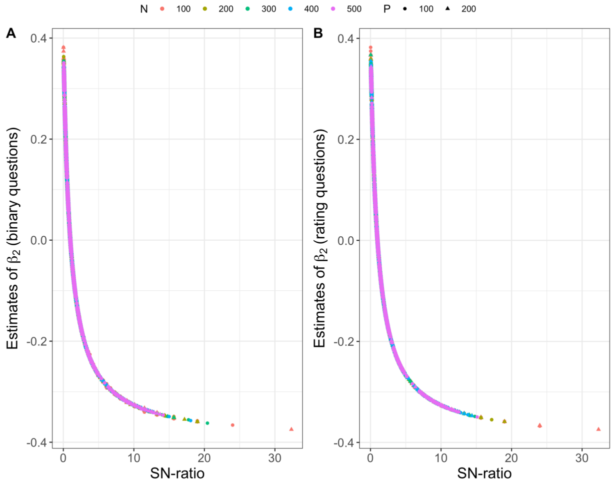

Figure 3. Relationships among N, P, SN-ratio, and estimates of β_2 for two types of questions: (A) binary questions; (B) rating questions.

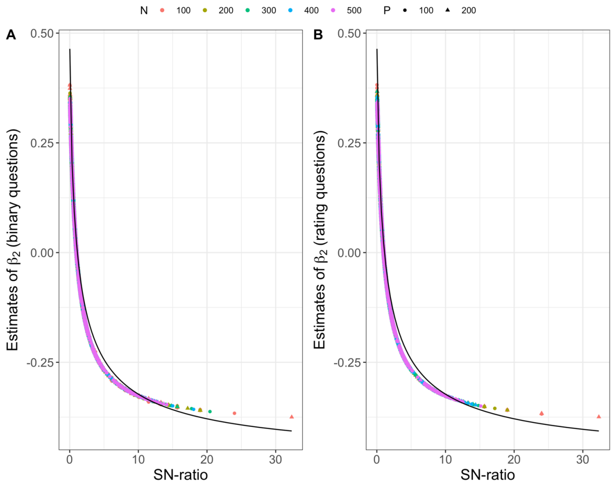

Figure 4. Simulated data (in color) and the fitted based on Eq. 8 (in black): (A) binary questions; (B) rating questions.

Estimates of β_3. For both binary and rating questions, it was found that the β_3 generally fluctuated around 0 (Figure 5 and Table 1) for different P and SN-ratio, and the estimates of β_3 would get closer to 0 with the increase of N. However, it was found that among all fitted models, the interaction effects between *d'* and *c* were significant in around 63% of all fitted models (i.e., 5680 of 9000 models for binary questions and 5608 of 9000 models for rating questions). To further reveal the contribution of these interaction terms in the model, using the same data, we fitted other 18,000 models, while without the interaction term between *d'* and *c*. It was found that all the models without the interaction term still had R^2 values of over 0.999. In other words, although the interaction effect between *d'* and *c* was significant in some cases, the contribution of this interaction effect to the prediction of PC was almost omittable. Hence, we set β_3 as 0.

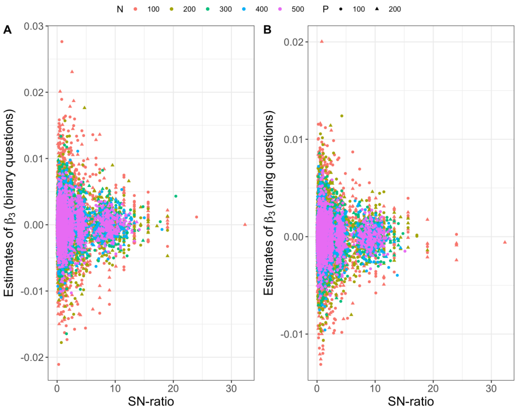

Figure 5. Relationship among N, P, SN-ratio, and estimates of β_3 for two types of questions: (A) binary questions; (B) rating questions.

The empirical equation. Based on the above results, the empirical equation describing the relationships among *d'*, *c*, and PC can be written as:

$$PC = 0.5 + 0.2 * d' + \beta_c * c \qquad (9)$$

$$\text{where, } \beta_c = \frac{e^{-0.73ln(SN-ratio)}}{1+e^{-0.73ln(SN-ratio)}} - 0.5 \qquad (10)$$

A Survey-based Case Study

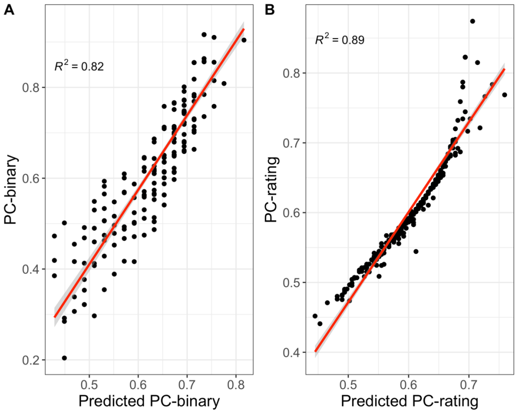

Figure 6. Comparisons between predicted PC from Eq. (9) and (10), and PC extracted from Survey Data: (A) binary questions; (B) rating questions.

To explore the accuracy of using *d'* and *c* to estimate PC based on the empirical equation we extracted from the Simulated Data, we compared the predicted PC and extracted for both binary answers and rating answers. As shown in Figure 6, two linear regression models were fitted. The R^2 for both types of questions were over 0.8, indicating high validity of our empirical equation.

DISCUSSION

Based on the data generated from a numerical simulation, this study explored the relationships between the sensitivity (*d'*) and response bias (*c*) calculated based on SDT and the proportion correctness calculated based on the methods adopted in previous research. Two types of questions (i.e., binary questions and rating questions) were considered when generating the Simulated Data for equation extraction. Using the Simulated Data, we quantified an empirical equation, in which the PC can be estimated based on *d'* and *c*. We further validated the extracted equation using the data collected from a real-world survey study. The PC predicted by the extracted empirical equation and the PC derived from participants' responses were compared. It was found that the extracted empirical equation reached satisfying performance in predicting PC for both binary (i.e., $R^2 = 0.82$) and rating questions (i.e., $R^2 = 0.89$). Thus, this equation may facilitate future meta-analysis and across-study comparisons when different mental model metrics were used in different studies.

Some interesting characteristics of the extracted empirical equation should be noted here. First, with $\beta_1 = 0.2$, it is straightforward to understand that higher *d'* would lead to higher PC-based score. At the same time, the effect of response bias on PC (i.e., the value of β_c) was influenced by the SN-ratio. More specifically, the β_c is positive when the SN-ratio < 1; the β_c becomes negative when the SN-ratio > 1; and the β_c become 0 when the SN-ratio = 1. In other words, for two individuals who have similar level of understanding of a system (d'), when answering a questionnaire that include fewer signals than noises (i.e., SN-ratio < 1), the participant who is more risky (i.e., smaller c, tends to say yes) would reach lower accuracy compared to the one who is more conservative (i.e., larger c, tends to say no). In contrast, when the number of signals was larger than the number of noises (i.e., SN-ratio > 1), the risker individuals would reach higher accuracy when answering the questions. When the number of signals was equal to the number of noises (i.e., SN-ratio = 1), the conserveness (or riskiness) of individuals would have no impact on the accuracy they obtain. Further, it should be noted that, the more extreme the SN-ratio, the larger the influence of *c* (i.e., respondents' bias) on the PC-based score.

Another interesting implication of the equation is that the PC would only be predicted by *d'* when $\beta_c * c$ was 0. This can happen in two possible scenarios: (1) when $c = 0$, that is, the respondent has no bias towards saying yes or no; (2) $\beta_c = 0$, that is, the SN-ratio in the task is 1 (i.e., there is equal number of signals and noises). The first scenario is straightforward, as *c* = 0 means the respondent has no bias. The second scenario can provide some implications for future studies. If we need to obtain an objective understanding of respondents' knowledge of a system and nullify the influence of their bias, we should set the SN-ratio in a designed questionnaire to be 1. At the same time, for future meta-analysis, the cross-study comparisons of PC-based scores can be performed only when the SN-ratio in the two studies are the same. Further, when the PC was only predicted by *d'* (i.e., $\beta_c * c = 0$), the empirical equation is simplified to $PC = 0.5 + 0.2 * d'$. Thus, the higher the *d'* (i.e., better capability to distinguish signals and noises), the higher the PC.

The intercept (β_0) is 0.5 in the empirical equation, which is also straightforward. When a respondent knows nothing about a system (i.e., *d'* = 0) and has no bias towards saying yes or no (i.e., *c* = 0), the answers from the respondent would be totally by chance (i.e., 50% probability of answering the questions correctly, leading a PC-based score of 0.5).

In summary, we extracted an empirical equation to quantify how *d'* and *c* can be used to estimate PC. The implications obtained from the equation can provide insights on the design of the mental-model-related questionnaires. The equation can also facilitate future cross-study comparisons of mental models. However, it should also be noted that the empirical equation is not validated through mathematical proof, as when calculating *d'* and *c*, the probabilities have to be transferred from z-scores using the cumulative distribution function of normal distribution, which, unfortunately has no closed-form solutions. To this end, future research may further validate the effectiveness of the empirical equation extracted in our study using more field data.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (No. 52202425), and in part by Guangzhou Municipal Science and Technology Project (No. 2023A03J0011), Guangzhou Science and Technology Program City-University Joint Funding Project (No. 2023A03J0001) and Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYB-2020083).

REFERENCES

Beggiato, M., & Krems, J. F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation Research Part F: Traffic Psychology and Behaviour*, *18*, 47–57.

Beggiato, M., Pereira, M., Petzoldt, T., & Krems, J. (2015). Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *35*, 75–84.

DeGuzman, C. A., & Donmez, B. (2021a). Drivers Still Have Limited Knowledge About Adaptive Cruise Control Even When They Own the System. *Transportation Research Record: Journal of the Transportation Research Board*, *2675*(10), 328–339.

DeGuzman, C. A., & Donmez, B. (2021b). Knowledge of and trust in advanced driver assistance systems. *Accident Analysis & Prevention*, *156*, 106121.

Dickie, D. A., & Boyle, L. N. (2009). Drivers' understanding of adaptive cruise control limitations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *53*(23), 1806–1810. SAGE Publications Sage CA: Los Angeles, CA.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Larsson, A. F. L. (2012). Driver usage and understanding of adaptive cruise control. *Applied Ergonomics*, *43*(3), 501–506.

McDonald, A., Carney, C., & McGehee, D. V. (2018). *Vehicle owners' experiences with and reactions to advanced driver assistance systems*.

National Highway Traffic Safety Administration. (2022). *Summary Report: Standing General Order on Crash Reporting for Level 2 Advanced Driver Assistance Systems*. Retrieved from https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-06/ADAS-L2-SGO-Report-June-2022.pdf

Parasuraman, R., Masalonis, A. J., & Hancock, P. A. (2000). Fuzzy signal detection theory: Basic postulates and formulas for analyzing human and machine performance. *Human Factors*, *42*(4), 636–659.

R Core Development Team, R Core Team, & Team, R. D. C. (2019). R: A Language and Environment for Statistical Computing. *Vienna, Austria*.

Rossi, R., Gastaldi, M., Biondi, F., Orsini, F., De Cet, G., & Mulatti, C. (2020). *A Driving Simulator Study Exploring the Effect of Different Mental Models on ADAS System Effectiveness*.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.