Part VII

Free Talk Session 4

JOHN SWETS, ROC CURVES, AND FIFTY YEARS OF SIGNAL DETECTION THEORY AND PSYCHOPHYSICS

Roger D. Adams[1], Gordon Waddington[1], Nili Steinberg[2], and Jia Han[3]
[1]*Faculty of Health, University of Canberra, Bruce, ACT 2617, Australia*
[2] *Wingate College of Physical Education and Sport Sciences, Wingate Institute, Netanya, Israel*
[3]*Shanghai University of Sport, Yangpu Qu, Shanghai Shi, China*
<roger.adams@canberra.edu.au; gordon.waddington@canberra.edu.au; knopp@wincol.ac.il; jia.han@canberra.edu.au>

Abstract

One hundred years after Fechner published the Elements of Psychophysics, the Psychometric Society invited involved researchers to contribute commentary papers to a special volume of Psychometrika. In the first article, Edwin Boring argues that Fechner's contribution in founding psychophysics was 'inadvertent'. Two of the papers, by Stanley S Stevens and John A Swets, represent important arms of the psychophysical enterprise, because they define the interests of their authors, in scaling through magnitude estimation and in discrimination through ROC curve analysis, respectively. The death of John Swets in 2016 marks an endpoint that warrants consideration of the half-century of work by the co-author of 'Signal Detection Theory and Psychophysics'. In a current application of ROC analysis in movement psychophysics, the Area Under the Curve is used both as an index of sensory acuity and as an indicator of injury vulnerability, where Youden's Index provides a rule for making decisions about cutoffs.

In 1961, the journal Psychometrika published papers on psychophysics that had been invited presentations for their 1960 Anniversary meeting, held one hundred years after the publication of Fechner's 'Elements of Psychophysics', in 1860. Three of the invitees were Edwin Boring (1886–1968), Stanley S Stevens (1906–1973), and John A Swets (1928–2016). In his article, Boring depicted Fechner as a cavalry commander who attacked the ramparts of materialism, then was inadvertently decorated for measuring sensation. He observed that Stevens' students heard about Fechner, and seldom missed celebrating October 22nd. Stevens questioned how Fechner's concept of error, the jnd, could have become a yardstick for measuring sensation, based as it was on Fechner's erroneous inspiration while lying in his bed, and concluded with his belief in the usefulness of ratio scales of sensation. Swets, however, suggested that Fechner would have welcomed Signal Detection Theory, recognizing in it the ideas about statistical decisions. In his conclusion, Swets (1961) suggested that Signal Detection Theory techniques employed with simple signals, such as operating characteristics, might be applied to more complex areas of research.

Twelve years later, in a paper titled The Relative Operating Characteristic in Psychology, Swets (1973) wrote that the ROC was a technique that effectively isolated the effects of observer response bias in the study of discrimination behavior, and gave a measure of discrimination that was independent of the location of the decision criterion. He noted that although the first appearance of ROC in the literature was in the Tanner and Swets (1954) paper, the Peterson and Birdsall (1953) technical report 'showed us how to plot the data'. In the title of his 1973 paper Swets used the term Relative Operating

Characteristic, though noting that originally, in the detection context, the R stood for Receiver. In his final major paper, Swets, Dawes and Monahan (2000), however, the R in the acronym is again Receiver. In his autobiography, Swets (2010) says 'the new terminology just didn't catch on'. Swets (1973) pointed out that if a rating scale is used to obtain better definition of the ROC curve, the cumulative technique employed resulted in a monotonic, increasing curve. Stanislaw and Todorov (1999) state that rating tasks, with r responses, are primarily used to measure sensitivity, and that the area under the ROC curve can be plotted from the r-1 points arising, giving a measure of sensitivity unaffected by response bias. A straightforward interpretation of the ROC area is that it is the proportion of times a subject would identify a signal, if signal and noise were presented simultaneously (Green & Swets, 1966). The area under a ROC curve (AUC) created from confidence ratings can be estimated by application of the trapezoidal rule (Brown, 1974). Because the area of a trapezoid is the multiple of the average length of the two parallel sides by the distance between them, the AUC can be calculated as a sum of a set of trapezoidal areas.

Finally, Swets, Dawes and Monahan (2000) outlined the value of using a non-parametric Signal Detection Theory approach to the problem of obtaining a discriminability measure, and show that not only is the AUC a discrimination measure, but that the ROC curve can be used to determine the best cutoff on a continuous variable to predict a binary state. It is this second application of ROC curves that is the focus of the present study.

When we began to use ROC curves to obtain a discrimination sensitivity measure for the extent of movement at joints (proprioception), we employed five stimuli and five responses, so that the resulting absolute judgement task, by giving the subject a continuous set of numbers to respond with, could be considered to also be a rating scale task. Previous findings have shown the obtained ankle movement discrimination scores, representing proprioceptive sensitivity, to be related to level of athletic achievement (Han et al, 2015), and to be improved by the use of textured insoles (Steinberg et al, 2015) and by training on an unstable balance board (Waddington & Adams, 2004).

The aim of this study was to examine the consequences of a recent ankle injury and to determine the score cutoff for recommending rehabilitation training.

Method

Participants

Forty-two full-time elite classical ballet dancers were in the current study. There were 27 dancers, 13 to 16 years old, and 15 dancers, 16–19 years old. All dancers or their parents provided written informed consent for participation. A survey about any ankle injury in the previous two years was completed prior to participation.

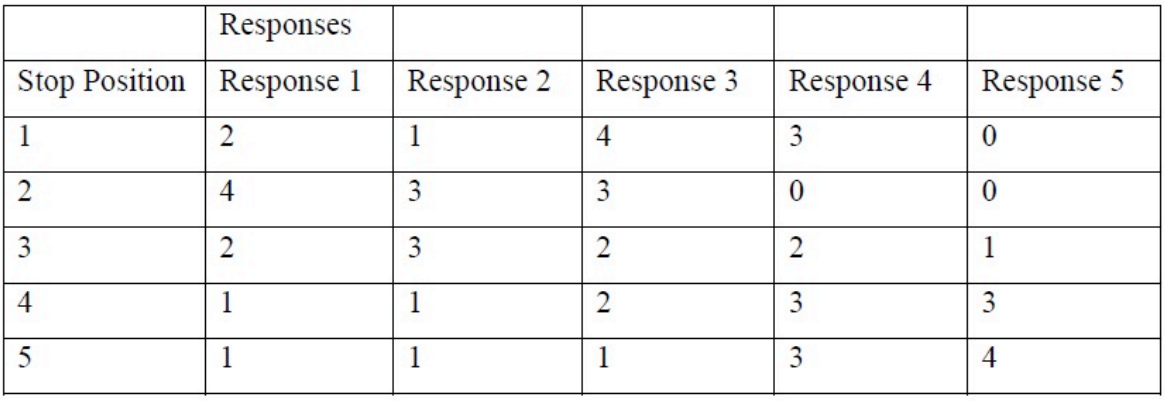
Equipment

Dancers were tested on their non-dominant weight-bearing leg for sensitivity to the extent of active ankle inversion movement on the Active Movement Extent Discrimination Apparatus (AMEDA)

The AMEDA device generates inversion stimuli with small differences in the extent of ankle inversion (Figure 1). With feet shoulder-width apart, the participants stood in a relaxed posture on the AMEDA platform, 50 cm above the floor, with the test foot centred over the axis of the movable base plate. Participants were first given a series of trial movements to perform, to familiarize them with the feel of the five different ankle inversion angles. Each participant performed a total of 15 movements in order to position 1 through 5 in sequence, three times. Participants then undertook 50 non-feedback trials, where they had to respond to the felt ankle inversion position. The trials were in a random sequence of 50, with 10 at each of the five different movement displacements. The stop positions of the stepper motor, operating a moving shaft vertically under the plate, were computer-determined, and ranged between 8 and 12 degrees of inversion, in one degree steps. During the 50 trials, the participants were asked to move down and back to horizontal at a steady pace. After each movement the participant was asked to respond with the number (level) that described the angle at which the device stopped. Software in a laptop computer was used to control a spinning shaft and set the vertical stops for each movement. Thereafter, ROC analysis was used to generate discrimination scores representing each participant's sensitivity to small differences in the extent of ankle inversion.

Data analysis

Data were cast into 5×5 confusion matrices representing the responses made to movements to each of the five stop positions, giving ten trials per row. An example data set is reproduced in the table below.



	Responses				
Stop Position	Response 1	Response 2	Response 3	Response 4	Response 5
1	2	1	4	3	0
2	4	3	3	0	0
3	2	3	2	2	1
4	1	1	2	3	3
5	1	1	1	3	4

The rows were then selected as adjacent pairs, the corresponding ROC curves

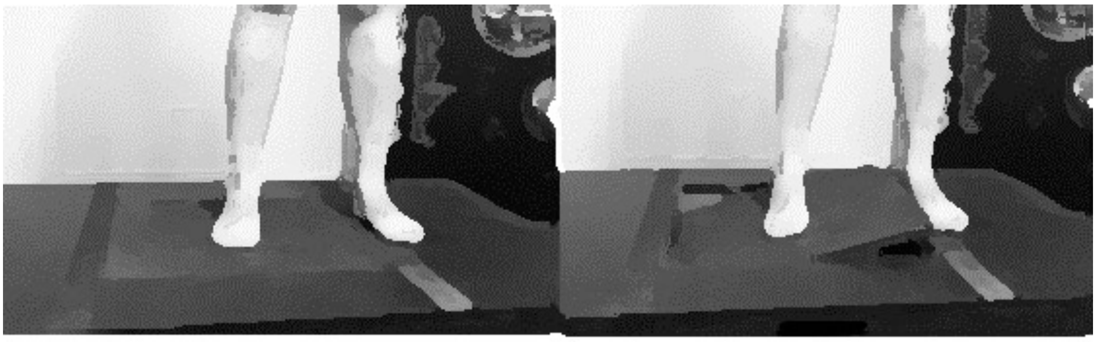


Fig. 1. AMEDA with plate at horizontal, and with ankle inversion to a stop at 10 degrees.

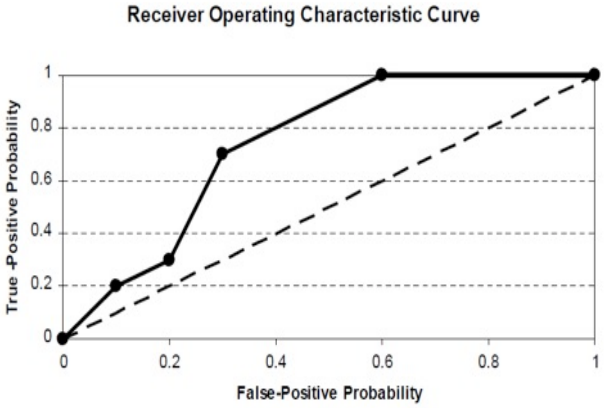


Fig. 2. ROC curve for an adjacent pair of ankle inversion extents. Dropping lines from the points on the curve to the abscissa enables computation of the area in each of the five resulting trapezoids, which when summed and converted to a proportion of the total area give the Area Under the Curve.

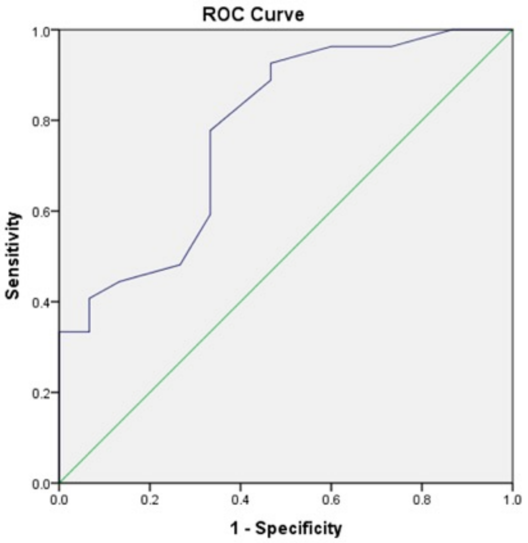


Fig. 3. ROC curve for ankle AMEDA score and Injury/No Injury, with 42 AUC values as the points on the curve, and injury as a 'hit'.

drawn, and the AUC calculated using the trapezoidal rule (Brown, 1974).

Results

Once the four AUC values for each participant were obtained, a mean was calculated to give a single score for ability to discriminate between ankle inversion movements separated by one degree, in the 8 to 12 degrees of ankle inversion range. Of the 42 dancers, 15 reported that they had sprained an ankle in the past two years, and 27 reported no injury. For the group with no ankle injury in the preceding two years, the mean AUC score was M = 0.687 (95% CI 0.668–0.707) whereas for the group with ankle injury the values were M = 0.635 (95% CI 0.608–0.662) and these means were significantly different ($p < 0.01$).

Thereafter, the mean AUC scores as a continuous variable and injury status as a binary variable were entered into ROC analysis. The resulting AUC was 0.778, p = 0.003, 95% CI = 0.629–0.927. Next, Youden's Index was calculated from the Sensitivity and Specificity values, and the AUC value that corresponded to the local maximum identified.

Youden's Index = Sensitivity - (1 - Specificity)

Test Variable: AMEDA_AUC score

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity	Y.I.
.000000	1.000	1.000	0.000
.585000	1.000	.867	0.133
.605000	.963	.733	0.230
.615000	.963	.600	0.363
.625000	.926	.467	0.459
.635000	.889	.467	0.422
.645000	.778	.333	0.445
.655000	.667	.333	0.334
.665000	.593	.333	0.260
.675000	.481	.267	0.214
.685000	.444	.133	0.311
.695000	.407	.067	0.340
.705000	.333	.067	0.266
.715000	.333	.000	0.333
.725000	.259	.000	0.259
.745000	.148	.000	0.148
.765000	.111	.000	0.111
.775000	.074	.000	0.074
.785000	.037	.000	0.037
1.000000	.000	.000	0.000

Discussion

The ROC analysis of the AMEDA ankle inversion discrimination AUC scores of dancers with and without ankle injury in the past two years showed that the AUC score was a significant discriminator between the groups, although at 0.78 it was below the level of 0.8 that Swets gives as the AUC for a test that is a 'good discriminator'. The local maximum in Youden's Index of 0.459, at a mean AUC score of 0.625, provides a cutoff AUC value for differentiating dancers with a previous ankle injury from those with no injury, and by inference a level of ankle inversion discrimination from testing on the AMEDA that warrants ankle rehabilitative work for dancers performing at this level.

Conclusion

The method of ROC curve analysis that John Swets worked to develop provides, in the AUC, a robust measure of discrimination accuracy that has proved to be both sensitive and effective in studies of discrimination of the extent of movements made at different body joints. In particular, at the ankle the measure reflects differences between injured and non-injured ankles, and ankle proprioception differences between athletes with different levels of achievement.

References

Boring, E. G. (1961) Fechner: Inadvertent founder of psychophysics. *Psychometrika,* 26, 3-8.

Brown, J (1974) Recognition assessed by rating and ranking. *British Journal of Psychology,* 65(1), 13-22.

Green, D. M., & Swets, J.A. (1966) *Signal Detection Theory and Psychophysics.* New York: Wiley.

Han, J., Waddington, G., Anson, J., Adams, R. (2015) Level of competitive success achieved by elite athletes and multi-joint proprioceptive ability. *Journal of Science and Medicine in Sport,* 18(1), 77-81.

Peterson, W.W. & Birdsall, T.G. (1953) The Theory of Signal Detectability. *Technical report* No. 13, Electronic Defense Group, University of Michigan, Ann Arbor.

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

Steinberg, N., Waddington, G., Adams, R., Karin, J., Begg, R., Tirosh, O. (2015) Can textured insoles improve ankle proprioception and performance in dancers? *Journal of Sports Sciences,* 34(15), pp. 1430-1437.

Stevens, S. S. (1961) Toward a resolution of the Fechner-Thurstone legacy. *Psychometrika,* 26, 35-47.

Swets, J.A. (1961) Detection theory and psychophysics: A review. *Psychometrika,* 26(1), 49-63.

Swets, J. A. (2010). *Tulips to Thresholds: Counterpoint Careers of the Author and Signal Detection Theory.* Los Altos Hills: Peninsula Publishing.

Swets, J. A. (1973) The Relative Operating Characteristic in Psychology. *Science,* 182, 990-1000.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better decisions through science. *Scientific American,* October, 82-87.

Tanner W.P. & Swets, J.A. (1954) A decision-making theory of visual detection. *Psychological Review,* 61(6), 401-409.

Waddington, G. & Adams, R. (2004) The effect of a 5-week wobble-board exercise intervention on ability to discriminate different degrees of ankle inversion, barefoot and wearing shoes: A study in healthy elderly. *Journal of the American Geriatrics Society,* 52, 573-576.

GRAY COLOR CATEGORICAL PERCEPTION: ASYMPTOTICS OF PSYCHOMETRIC FUNCTION AND MEAN DECISION TIME

Ren Namae, Marie Watanabe, and Ihor Lubashevsky
University of Aizu
Ikki-machi, Aizu-Wakamatsu, Fukushima 965-8580, Japan
IL: `<i-lubash@u-aizu.ac.jp>`

Abstract

We summarize the results of our experiments on categorical perception with respect to gray color categorization into two and three classes. Namely, the subjects were instructed to categorize shades of gray (generated in a random sequence) making selection between light-gray and dark-gray (first set of experiments) or between light-gray, gray, and dark-gray (second set of experiments). The collected data are analyzed employing (*i*) *the asymptotics of the constructed psychometric functions and* (*ii*) *the mean decision time in categorizing a given gray shade. A plausible macro-level mechanism governing gray color categorization is discussed.*

The notion of categorical perception generally describes situations when we perceive our world in terms of discrete categories emerged previously during our communication with the social environment. Our perception is warped such that difference between objects belonging to different categories is accentuated and, in opposite, difference between objects falling into one category is deemphasized (for a review see, e.g., Harnad, 2005; Goldstone and Hendrickson, 2010). As far as color categorization is concerned, in spite of a vast amount of literature about various aspects of color categorization (e.g., Harnad, 2005) the basic mechanisms governing these processes remain up to now a challenging problem.

In the present work we summarize our previous experiments on gray color categorization and shape recognition (Lubashevsky and Watanabe, 2016; Namae et al., 2017) as well as argue for the existence of a certain emergent mechanism of decision making under uncertainty governing the categorical perception at the macro-level rather than the level of particular neurophysiological processes. The pivot point of our experiments is the analysis of (*i*) the *asymptotic behavior* of the corresponding psychometric functions and (*ii*) the mean decision time in classifying a given shade of gray (shade number).

Experimental Setup and Data Processing

On Lenovo LI2221s Monitor (47.7 × 26.8 cm screen) a computer under the operating system Windows 10 visualizes a window of size of 17 × 16 cm with a square $\mathbb{S}$ of size of 11 × 11 cm placed at it center. Color inside this square is changed during experiments; the remaining window part is filled with a neutral gray, namely, RGB(240,240,240). The brightness and contrast of the screen was set equal to 70% and 60%, respectively. To get subject's response to a visualized color we used a standard game joystick. The same computer was used for all the experiments. Integrally eight subjects (five male and three female students) were involved in these experiments.

Each trial of shade categorization is implemented as follows. A random integer $I \in [0, 255]$ is generated uniformly and the area $\mathbb{S}$ is filled with the gray color $G(I) = RGB(I, I, I)$. Then, in one set of experiments, subjects have to classify the visualized gray color $G(I)$ into two categories: *light gray* and *dark gray*, within the other set of experiments into three categories: *light gray*, *gray*, and *dark gray*.

A subject's choice is recorded via pressing the corresponding buttons of a standard joystick. Then a mosaic pattern of various shades of gray is visualized for 0.5 s. This mosaic pattern is used to depress a possible interference between color perception in successive trials that can be caused by human iconic memory. Figure 1 illustrates this. After that a new number I is generated uniformly and the next trial starts. Each data point comprises three quantities (I, J, Δ): the currently generated index I of gray shade, the index J of currently chosen category, and the decision time Δ, i.e., the time interval between visualizing the gray color $G(I)$ and the instant when a subject presses the corresponding joystick button.

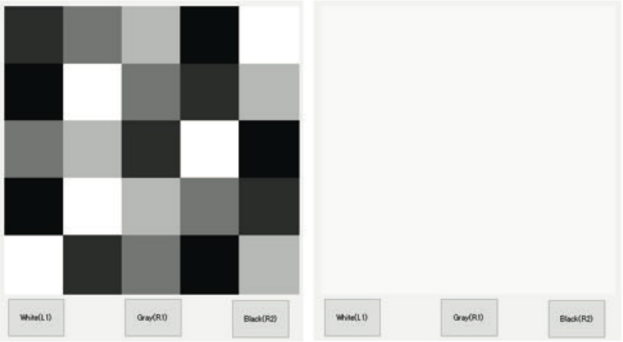


Fig. 1. Color generator

Experiments for each subject spanned over 4 days. We have recorded 2000 data points per day and the total amount of data points for one subject individually is $M = 8000$. This amount of data is crucial because, on one hand, only the asymptotic behavior of psychometric functions bears the information about possible mechanisms governing subject's decision-making. On the other hand, the relative volume of collected data related to this asymptotics is rather small.

The collected data have been used for constructing the following two functions. The first one is the psychometric function for a given gray color category J_α, e.g., the "light-gray" category, i.e., the probability $P_w(I)$ of classifying a given shade I as "light-gray"

$$P_\alpha(I) = \left[\sum_{k=1}^{M} \delta(I, I_k)\delta(J_\alpha, J_k)\right] \cdot \left[\sum_{k=1}^{M} \delta(I, I_k)\right]^{-1}, \tag{1}$$

where k is the index of recorded data point and $\delta(i, j)$ is the Kronecker delta. The second function $T(I)$ is the mean decision time of choosing any one of the possible categories for a given shade of gray:

$$T(I) = \left[\sum_{k=1}^{M} \delta(I, I_k)\Delta_k\right] \cdot \left[\sum_{k=1}^{M} \delta(I, I_k)\right]^{-1}. \tag{2}$$

The two functions are used to single out the characteristic properties of plausible mechanisms governing the analyzed categorical perception.

Results and Discussion

Figures 2 and 3 illustrate the obtained results. The logarithmic scale of the P(i.e., y)-axis is used to visualize the asymptotic behavior of the constructed psychometric functions $\{P_\alpha(I)\}$ *vs* the shade number I of gray color. The frames are arranged in such a way that the psychometric function plot and the mean decision time plot combined together for each subject individually be placed one above the other. This arrangement makes it clear that the peaks in the dependence of mean decision time on the shade index correspond to the regions where the uncertainty in gray color categorization is most pronounced.

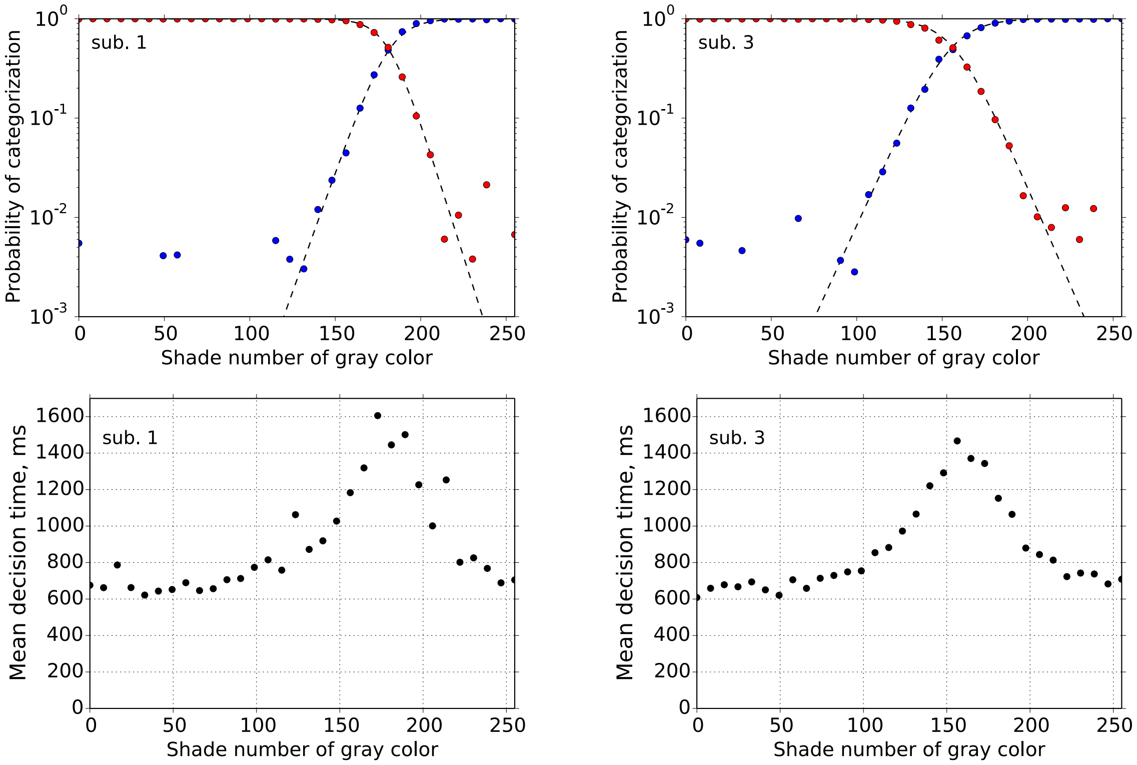


Fig. 2. Psychometric functions $P\alpha(I)$ (or $1 - P\alpha$) and the mean decision time $T(I)$ vs the shade of gray G(I) for two subjects involved in the two-categories-choice experiments.

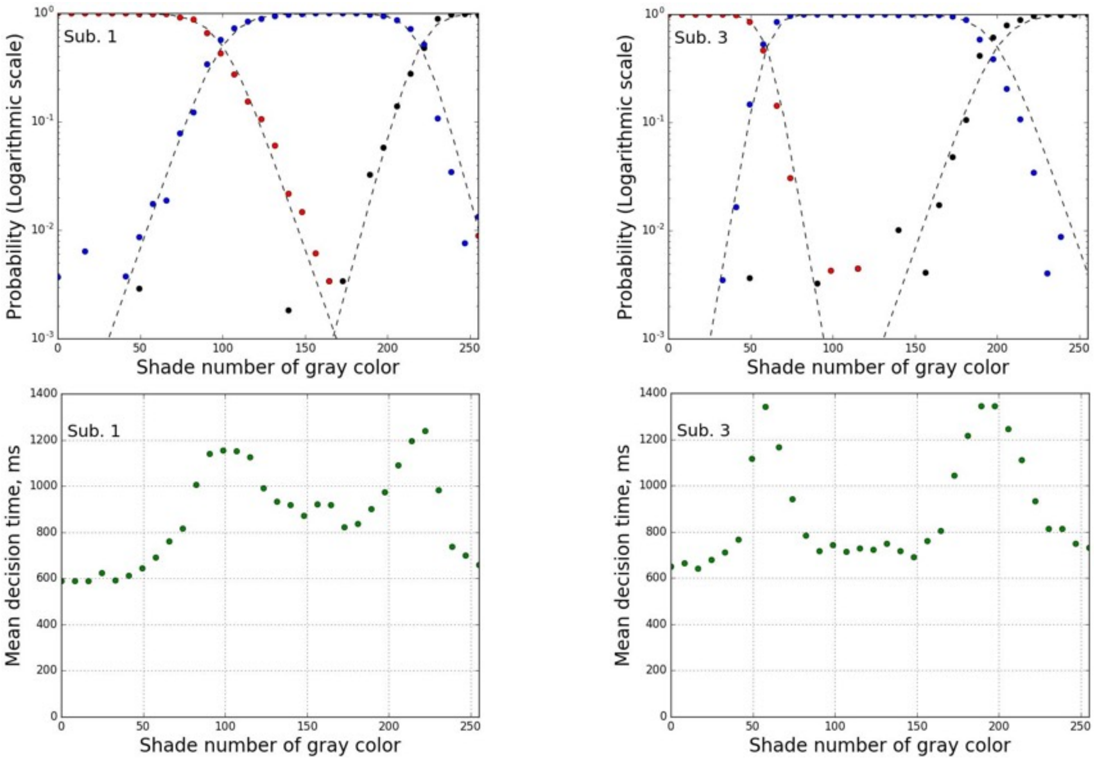


Fig. 3. Psychometric functions $P\alpha(I)$ (or $1-P\alpha$) and the mean decision time $T(I)$ vs the shade of gray G(I) for two subjects involved in the three-categories-choice experiments.

As seen, there are three common features exhibited by the data collected for all the subjects in the presented experiments.

First, it is the exponential form of their asymptotic behavior. In the shown log-normal plots dashed lines represent the fitting logistic-type functions specified by the fitting function:

$$P_\alpha(I) \approx \frac{1}{2}\left\{1 + \tanh\left[\pm\frac{(I - I_{m:\alpha})}{\delta I_\alpha}\right]\right\}, \tag{3}$$

where $I_{m:\alpha}$ is the center point of the crossover region and the value δI_α characterizes its thickness. The used parameters of this approximation in fitting the experimental data are individual for each subject.

These results allow us to suppose that the mechanism governing the uncertainty in subject's decision-making in categorizing the gray shades admits interpretation as the potential type mechanism. It means that the subject's decision-making is treated as a random process ζ (where ζ is the state variable of some system) in a potential field $U(\zeta, I)$ depending on the shade index I as a parameter. In this case the subject's choice of a category α at trial k for a given shade index I is described by the steady-state probability of finding the system inside the potential well located at the point ζ_α:

$$P_\alpha \approx \exp\{-U_\alpha(I)\}\left[\sum_{\alpha'} \exp\{-U_{\alpha'}\}\right]^{-1} \tag{4}$$

Here $U_\alpha(I) := U(\zeta, I)|_{\zeta=\zeta_\alpha}$ and we have assumed all the potential wells to be rather deep to ignore the contribution of other system states to probability (4). In the general case when for a certain value of the grade index I the contributions of two potential wells become comparable, the other potential wells remain ignorable. So within the linear approximation of the potential dependence on the parameter I

$$U(\zeta, I + \delta I) \approx U(\zeta, I) + \frac{\partial U(\zeta, I)}{\partial I} \cdot \delta I$$

expression (4) leads to formula (3). Other models like detecting noisy signals give another type asymptotics of psychometric functions.

Second, the presented plots show how the mean time required for the subjects to make decision about classifying a current shade of gray changes with the shade number I. For the values of I corresponding to the crossover regions the patterns $T(I)$ contain peaks. In particular, for all the subjects the mean decision time is found to be about 600–700 ms outside these peaks, which can be regarded as the upper boundary T_p of the human response delay time controlled by physiological processes of recognizing threshold events within their unpredictable appearance. At least, it is the upper boundary of visual time intervals presenting timescales relevant to natural behavior, see, e.g., Mayo and Sommer (2013) and references therein.

In the crossover regions the uncertainly in subject's category choice is most essential because for such values of shade number choosing one of two categories becomes equiprobable. Here the peak values T_m of the mean decision time exceed T_p substantially, e.g., for the shown data $T_m \approx 1.4$ s. It should be also noted that a similar dependence of the decision time was found in the speech recognition Bidelman et al. (2013), however, the time delay attained in the peak maximum does not exceed 600 ms. In our data the mean time of decision-making includes also a time interval between making decision and pressing the corresponding button. However, if the time delay is deducted from the measured time data, the found peak will be even more pronounced.

The appearance of these peaks at the $T(I)$-patterns can be explained by turning to the concept of dual system of decision-making. This concept accepts the existence of the automatic and intentional systems contributing simultaneously to human response and being in a continuous interplay with each other. The former—automatic system—is reflexive, fast, affective, associative, and primitive. The latter—intentional system—is deliberative, controlled, slow, cognitive, propositional, and more uniquely human. Besides, there are accounts assuming the dual-processes to arise parallel and compete with each other. However, there are also arguments against the dual system of decision-making; for a review and discussion of the evidence supporting both sides of debate a reader may be referred, e.g., to Rustichini (2008); Evans (2008, 2011). The found peaks of the mean decision time dependence on the shade number argue for that the two cognitive systems do exist and are comparable in their influence on the categorical perception. Indeed, the characteristic time scale of decision-making in categorical perception depends substantially on the uncertainty in the subject's choice. It argues that conscious (mental) processes should be involved in decision-making when the choice of the appropriate color category is not evident.

Third, in categorizing any given shade of gray the subjects made choice only between two classes. Indeed, in the crossover regions the data points correspond to the subject's choice either between "light-gray" and "gray" or between "gray" and "dark-gray." For example, the crossover region of the choice between "light-gray" and "gray" practically does not contain the data points matching the "dark-gray" category.

Conclusion

The reported results argue for the following hypothesis:

- Categorical perception, at least, of shades of gray, is governed by a potential mechanism of decision-making that can be treated as a random process in a potential field whose profile depends on a given shade number as a parameter. Within this approach the classification classes match the wells of the corresponding potential relief.

- The characteristic time scale of decision-making in categorizing the shades of gray, at least in the analyzed case, depends substantially on the uncertainty in classifying a given shade; the higher the uncertainty, the longer the decision time. The obtained data enable us to relate this effect with considerable contribution of mental processes to categorization. We regard this feature as a certain argument for the existence of the dual system of cognitive processes.

- When it is physically possible and different categories are independent, i.e., they are not mixed in the mind—which is the case in the gray color categorization—humans prefer to make choice between only two categories of classification for a given object or stimulus, whereas other categories are not taken into account. We call it the principle of pair-wise categorization.

References

Bidelman, G. M., Moreno, S. and Alain, C.: 2013, Tracing the emergence of categorical speech perception in the human auditory system, *NeuroImage* **79**, 201–212.

Evans, J.: 2008, Dual-processing accounts of reasoning, judgment, and social cognition, *Annual Review of Psychology* **59**, 255–278.

Evans, J. S.: 2011, Dual-process theories of reasoning: Contemporary issues and developmental applications, *Developmental Review* **31**(2–3), 86–102. Special Issue: Dual-Process Theories of Cognitive Development.

Goldstone, R. L. and Hendrickson, A. T.: 2010, Categorical perception, *Wiley Interdisciplinary Reviews: Cognitive Science* **1**(1), 69–78.

Harnad, S.: 2005, To Cognize is to Categorize: Cognition is Categorization, *in* H. Cohen and C. Lefebvre (eds), *Handbook of Categorization in Cognitive Science*, Elsevier, Amsterdam, pp. 20–43.

Lubashevsky, I. and Watanabe, M.: 2016, Statistical Properties of Gray Color Categorization: Asymptotics of Psychometric Function, *Proceedings of the 47th ISCIE International Symposium on Stochastic Systems Theory and Its Applications Honolulu, Dec. 5-8, 2015*, Institute of Systems, Control and Information Engineers (ISCIE), Kyoto, pp. 41–49.

Mayo, J. P. and Sommer, M. A.: 2013, Neuronal correlates of visual time perception at brief timescales, *Proceedings of the National Academy of Sciences* **110**(4), 1506–1511.

Namae, R., Watanabe, M. and Lubashevsky, I.: 2017, Gray Color Multi-Categorical Perception: Asymptotics of Psychometric Function, *Proceedings of the 48th ISCIE International Symposium on Stochastic Systems Theory and Its Applications Fukuoka, Nov. 4-5, 2016*, Institute of Systems, Control and Information Engineers (ISCIE), Kyoto, pp. 76–80.

Rustichini, A.: 2008, Dual or unitary system? Two alternative models of decision making, *Cognitive, Affective, & Behavioral Neuroscience* **8**(4), 355–362.

SUBJECTIVE CONFIDENCE IN PERCEPTUAL CATEGORIZATION: ALTERING PERFORMANCE ASYMPTOTES INCREASES OVERCONFIDENCE

Jordan Richard Schoenherr and Guy Lacroix
Department of Psychology, Carleton University
1125 Colonel By Drive, Ottawa, ON K1S5B6 Canada
<Jordan.Schoenherr@Carleton.ca, Guy.Lacroix@Carleton.ca>

Abstract

Categorization represents the coordination of perceptual discrimination and memory processes. Contemporary models have acknowledged the conjoint contributions of implicit and explicit processes, notably in COVIS (competition between verbal and implicit systems). COVIS assumes that an explicit system engages in hypothesis-testing, dominating the early stages of learning. Over time an implicit system that uses feedback to engage in procedural learning, dominating later stages of learning. Studies supporting this theory have focused on categorization responses and concurrent tasks. Using the randomization technique, we presented participants with Gabor patches from two contrasting categories. We varied the extent to which exemplars from the categories overlapped, creating a performance asymptote (65%) allowing us to examine overconfidence. Participants required little training to reach this performance asymptote Importantly, in contrast to conditions with higher performance asymptotes (85%), we observed much greater overconfidence. This suggests that confidence reports are not solely or primarily determined by accumulated evidence.

The acquisition, development, and use of categories have been an enduring topic of study across psychology. These studies have yielded the progressively clearer understanding that category learning relies on two interactive, yet dissociable information processing systems. The first is a fast-learning, resource-limited explicit hypothesis-testing system that can learn in the absence of feedback. The second system is a slow-learning, high-capacity implicit procedural-learning system that is feedback-dependent (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Krushke, 1998; Nosofsky, Palmeri & McKinley, 1994). During learning, a representational shift occurs from early to late stages of learning: early stages of categorization are dominated by an explicit representational system whereas the later stages of learning are dominated by an implicit representational system (e.g., Ashby et al., 1998).

A straightforward implication of dual-process models of categorization such as COVIS (COmpetition between Verbal & Implicit Systems; Ashby et al., 1998) is that participants should have more explicit awareness of rule-based category structures relative to information-integration category structures. However, few studies have examined the role of subjective awareness (Paul et al., 2011) or have attempted to dissociate these learning systems with subjective measures in categorization (Schoenherr & Lacroix, 2012). The present study examines whether differences in subjective awareness are evidenced over the course of learning and whether different category structures are associated with greater reported levels of subjective awareness. Specifically, we employed post-decisional confidence reports to assess subjective confidence and compared these results with participants' accuracy (i.e., confidence calibration). By examining systematic variations in overconfidence, we provide evidence for changes in subjective awareness.

Competition between Verbal and Implicit Systems

COVIS assumes that participants make categorization decisions using a criterion or category boundary (e.g., Ashby & Gott, 1988). When participants are presented with a stimulus, they categorize it relative to the category boundary. With feedback, participants will adjust the location of the category boundary to maximize the separation between the categories. In the hypothesis-testing system, simple low-dimensional rules (e.g., stimulus size or frequency) are rapidly generated and tested. In the procedural-learning system, higher-dimensional rules (e.g., stimulus size and frequency) are slowly created and altered with feedback. Ashby et al. (1998) additionally assumed that the hypothesis-testing system and the procedural-learning system remain co-activated under most conditions. In this way, they compete for response selection with the hypothesis-testing system dominating early stages of training and the procedural-learning systems dominating later stages of training. While evidence supporting these results comes from an examination of response accuracy, we sought out a subjective measure of awareness that could be used throughout learning.

Confidence Reports and Confidence Processing

Subjective measures of performance such as confidence reports have a long history in psychophysical research (for a review, see Baranski & Petrusic, 1998). In most studies, participants provide a post-decisional confidence reports by assigning a subjective probability on a rating scale immediately after they have made a decision. For instance, in a two-alternative forced-choice (2AFC) task, participants might indicate that they were guessing (50%), that they had a reasonable level of confidence (e.g., 70%), or that they were certain in their response (100%). The degree of correspondence between a participant's mean accuracy when assigning a subjective probability to a response is referred to as subjective calibration (e.g., Baranski & Petrusic, 1994). Perfect calibration requires that the proportion correct (e.g., 0.7) and mean confidence are equivalent (e.g., 70%) whereas miscalibration such as overconfidence represents a bias. Studies of perceptual discrimination and general knowledge (e.g., Baranski & Petrusic, 1994) as well as memory (e.g., Koriat, 1993) have observed systematic deviations in the correspondence between task accuracy and subjective probabilities. These deviations can be attributed to differences in the operations supporting primary decision response selection and confidence processing.

COVIS contains specific assumptions concerning confidence. It assumes that confidence is obtained by a direct-scaling of this evidence obtained during the primary decision (Ashby et al., 1998). This conforms to SDT-based models of confidence process that claim that information obtained from the primary decision is the sole determinant of confidence report (e.g., Ferrel & McGooey, 1980). However, a number of studies have suggested that confidence reports are affected by sources of information other than that provided by the target stimulus (Busey, Tunnicliff, Loftus, & Loftus, 2000; Schoenherr, Leth-Steensen, & Petrusic, 2010) and require additional operations associated with increased decision response time (DRT) when they are reported (e.g., Baranski & Petrusic, 1998; Schoenherr, 2009). Direct-scaling models have difficulty accounting for such findings (cf. Pleskac & Busemeyer, 2010). Subjective awareness of the properties different category structures should influence participants confidence, with greater awareness of a categorical representation associated with greater confidence. On this account, overconfidence could suggest that while participants can access a representation of the category structure, this representation is not what is being used to categorize stimuli. We consider the basis for these predictions next.

Subjective Awareness in Category Learning

Schoenherr and Lacroix (2012) have explored the possibility of using confidence reports to dissociate category learning systems. They assumed that the degree of correspondence between measures of accuracy and confidence can be used to infer the accessibility of representations and the underlying architecture of categorization processes during different stages of learning. First, when a performance asymptote is used, confidence should reach an asymptote prior to accuracy given the flexibility of the hypothesis-testing system and should exhibit a more rapid learning rate. Second, they assumed that participants should exhibit overconfidence when the category structure is readily verbalizable. Third, the requirement of confidence should also increase DRT if it constitutes a secondary process. Moreover, if the hypothesis-testing system and confidence share the same basis, automaticity of responses should occur more rapidly in the rule-based condition relative to the information-integration condition.

Schoenherr and Lacroix (2012) obtained evidence across three experiments that supported dissociable category learning systems. Using an 85% performance asymptote (Experiments 1a and 1b) while also manipulating feedback in order to affect learning in the procedural-learning system but not in the hypothesis-testing system (Experiment 2), they observed increases in DRT when post-decisional confidence reports were required in comparison to a no confidence condition (Experiment 1a). They additionally observed increased overconfidence in intermediate phases of training for those participants learning a rule-based category structure relative to those who learned the information-integration category structure. The pattern of miscalibration suggested that the representation used to report subjective confidence were influenced by different sources of information. They interpreted the greater level of overconfidence as evidence for greater subjective awareness of rule-based representations that failed to contain information concerning the stimulus variability (i.e., they ignored exception exemplars).

In the present study, we sought to extend our previous findings. All manipulations were identical to those of Schoenherr and Lacroix (2012) except that a category overlap of 35% was used (See Ell & Ashby, 2006). The resulting performance asymptote of 65% was imposed to increase the proportion of negative feedback that participants received relative to Schoenherr and Lacroix (2012). We thus predicted that participants would exhibit greater overconfidence relative to our previous experiments if participants failed to account for the increase in negative feedback or exception exemplars.

Method

The stimuli consisted of Gabor patches varying in terms of spatial frequency and orientation. Replicating the method of earlier studies (e.g., Zeithamova & Maddox, 2007), 40 Gabor patches were created for each category for the training phase using the randomization technique by randomly sampling values from two normal distributions. Stimulus values were rescaled into stimulus dimensions with spatial frequency given by $f = .25 + (x_1/50)$ and orientation given by $o = x2(\pi/500)$. Using these values, stimuli were generated with the Psychophysics Toolbox (Brainard, 1997) using MATLAB R2008 (MathWorks, Matick, MA) with an 65% performance asymptote. After a categorization response was provided and a confidence report was obtained, a feedback signal was presented to indicate a participant's accuracy in completing the task. Stimuli were presented to participants using E-Prime experimental software on a Dell Dimension desktop PC.

Results and Discussion

We first examined whether any participants performed at or below chance using a one-sample t-test. It revealed that all participants' classification accuracy were above chance ($M = .660, SD = .071$) suggesting that participants had learned the category structures, $t(102) = 22.86, p < .001$. Thus, no participants were excluded in either the rule-based or information-integration conditions.

Proportion Correct

A 2 (Categorization Rule: rule-based vs. information-integration) × 2 (Confidence Condition: block-level confidence vs. trial-and-block confidence) × 12 (Experimental Blocks: 1-12) mixed-design ANOVA was performed on proportion correct, collapsing across response key. Replicating previous studies using similar category structures (Ell & Ashby, 2006), no interaction was observed between Categorization Rule and Experimental Block, $F(11, 924) = .92, p = .504$. Replicating the results of our previous experiments (Schoenherr & Lacroix, 2012), we observed a significant effect of Experimental Block, $F(11, 924) = 10.21, MSE = .01, p < .001, {\eta^2}_p = .11$, as well as the main effect of Categorization Rule, $F(1, 84) = 4.15, MSE = .06, p = .045, {\eta^2}_p = .05$. Neither the main effect of confidence condition, $F(1, 84) = .099, p = .754$, nor its interactions with Experimental Block or Categorization Rule were significant (all $Fs < .91$, all $ps > .50$).

Table 1. Averaged dependent variables for rule-based (1D) and information-integration (2D) category structures. Standard errors are presented in parentheses.

	$p(correct)$	M_{Conf}	Calibration	O/U
Rule-Based	.64 (.01)	81.96 (2.36)	.07 (.01)	.16 (.02)
Information-Integration	.67 (.01)	78.27 (2.17)	.05 (.01)	.09 (.02)

As in Schoenherr and Lacroix (2012), participants' accuracy increased over blocks of learning trials. In the present study, we observed that learning was slight better in the information-integration condition relative to the rule-based condition (see Table 1). However, the mean performance (M = .66) across both conditions was nearly identical to the desired performance asymptote (i.e., a learning criterion of .65).

Trial-Level Subjective Confidence Calibration

A 2 (Categorization Rule: rule-based vs. information-integration) × 4 (Experimental Phase: 1-4) repeated-measures ANOVA was conducted. An analysis of subjective calibration did not reveal any difference between Experimental Phases, $F(3, 132) = 2.088, p = .12$, and obtained only a marginal effect of Categorization Rule (see Table 1), $F(1, 44) = 2.83, MSE = .00, p = .10$. The interaction of Experimental Phase and Categorization Rule was also not found to be significant, $F(3, 132) = .428, p = .696$. Such

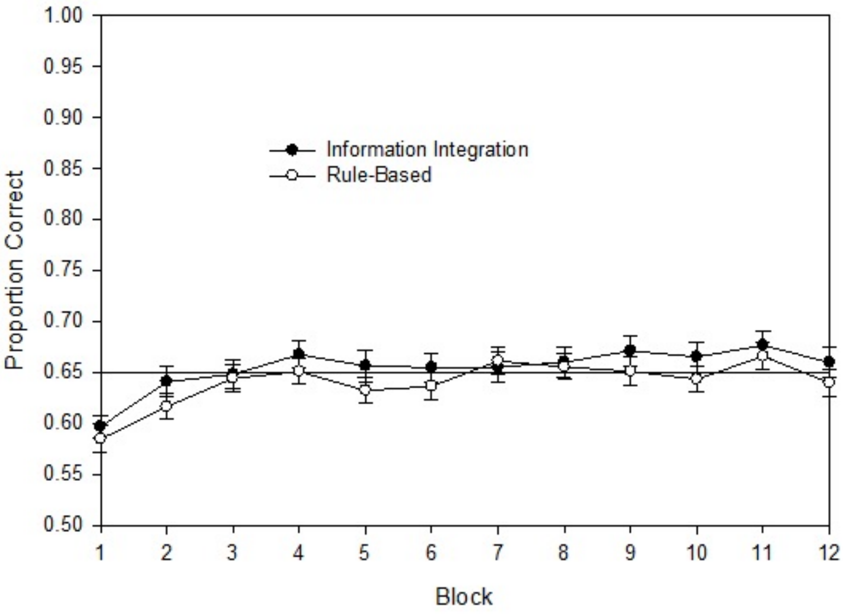


Fig. 1. Response accuracy with 65% performance asymptote. The error bars represent the standard error of the mean.

a finding is expected given that both accuracy and mean confidence reports reached an asymptote early in training. It suggests that, on a trial-to-trial basis, participants had a similar level of awareness of their performance although there is a trend toward improved calibration from the early phases of training to later experimental phases.

Trial-Level Overconfidence Bias

Unlike the calibration analysis, a significant difference in overconfidence bias was observed in the analysis of Experimental Phase, $F(3, 126) = 5.36, MSE = .01, p = .004, {\eta^2}_p = .11$. A significant effect of Categorization Rule was also observed, $F(1, 42) = 10.88, MSE = .02, p = .002, {\eta^2}_p = .21$. These findings suggest that participants' general awareness of the category structure differed between rule-based and information-integration category structures (see Table 1) as well as over experimental blocks of trials. As is evidenced in the nearly additive pattern in Figure 2, participants exhibited greater overconfidence in the rule-based condition relative to the information-integration condition. This finding would be expected if the participants in the rule-based condition created a representation that was more accessible to subjective awareness while failing to account for negative response feedback or exception exemplars that is used by an procedural-learning system. The pattern evidenced in Figure 2 also suggests that participants' overconfidence bias generally decreased over experimental blocks. This pattern suggests that the rapid generation of a representation within the hypothesis-testing system might lead participants to believe that their performance was in fact better than it was in early phases of the experiment (i.e., they "knew" what the categories looked like). With additional trials, participants' accuracy eventually caught up to their subjective confidence, thereby reducing overconfidence.

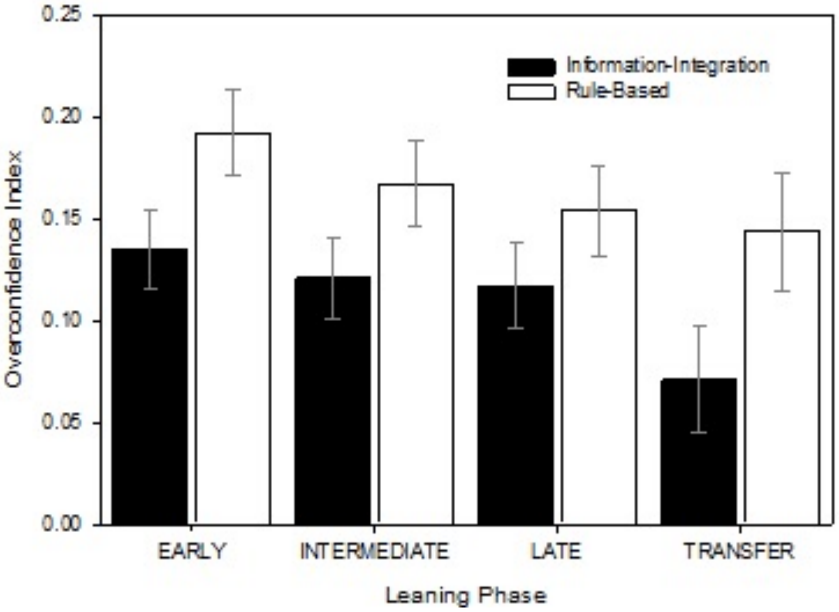


Fig. 2. Overconfidence bias for rule-based and information-integration category structures. Maximum overconfidence located at .35 (not plotted). The error bars represent the standard error of the mean. Phases reflected grouped sessions (e.g., Early: S1, S2; Transfer: S11, S12).

Conclusions

The results of our experiment replicated several earlier studies within the categorization and confidence processing literatures as well as our previous study (Schoenherr & Lacroix, 2012). Replicating Ell and Ashby (2006), while we found that learning was affected by categorization rule and expeirmental block, these two factors did not interact. Relative to our previous study (Schoenherr & Lacroix, 2012), we observed greater overconfidence across training and transfer. Thus, while the representation in the explicit learning system appears to be as accessible as in our previous studies thereby keeping confidence relative high, participants fail to account for their low performance and negative feedback. Thus, confidence reports appear to be determined by factors beyond those associated with the process of primary decision evidence accumulation and response accuracy. Taken together with our previous results, these findings suggest that multiple learning system acquire representations in a qualitatively different manner.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review,* 105, 442–481.

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences,* 9, 83–89.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *JEP: LMC,* 14, 33–53.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psycho-*

logical Review, 93, 154–179.

Baranski, J. V., & Petrusic, W. M. (1994). The Calibration and resolution of confidence in perceptual judgements. *Perception & Psychophysics,* 55, 412–428.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *JEP: HPP,* 24, 929–945.

Brainard, D. H. (1997). The Psychophysics Toolbox, *Spatial Vision,* 10, 433–436.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review,* 7, 26–48.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General,* 127, 107–140.

Ell, S. W.& Ashby, F. G. (2006). The effects of category overlap on information-integration and rule-based category learning. *Perception & Psychophysics,* 68, 1013–1026.

Ferrel, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behaviour and Human Performance,* 26, 32–53.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review,* 100, 609–639.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *JEP: LMC,* 29, 650–662.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition,* 22, 352–369.

Paul, E. J., Boomer, J., Smith, J. D., & Ashby, F. G. (2011). Information-integration category learning and the human uncertainty response. *Memory Cognition,* 39, 536–554.

Pleskac, T.J. and Busemeyer, J.R. (2010) Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review,* 117, 864–901.

Schoenherr, J. R. (2009). Mapping Internal Representations of Confidence onto Scales Varying in Range, Interval, and Number of Response Categories. Unpublished Manuscript.

Schoenherr, J. R., Leth-Steensen, C., & Petrusic, W. M. (2010). Selective attention and subjective confidence calibration. *Attention, Perception & Psychophysics,* 72, 353–368.

Zeithamova, D., & Maddox, W. T. (2007). The role of visuo-spatial and verbal working memory in perceptual category learning. *Memory & Cognition,* 35, 1380–1398.

CAPTURING DECISION CONFIDENCE THROUGH RESPONSE TRAJECTORIES AND WILLINGNESS TO GAMBLE

Arkady Zgonnikov[1,2], Aisling Kenny[1], Denis O'Hora[1], KongFatt Wong-Lin[3]
[1]*School of Psychology, National University of Ireland Galway, Galway, Ireland*
[2]*University of Aizu, Aizuwakamatsu, Fukushima, Japan*
[3]*Intelligent Systems Research Centre, Ulster University, Derry/Londonderry, UK*
<arkady.zgonnikov@gmail.com, denis.ohora@nuigalway.ie>

Abstract

We aimed to investigate whether action dynamics could be employed as an objective measure of decision certainty and the relationship between certainty and confidence. Twenty-eight participants were required to view a random dot kinematogram display and report the dominant dot direction by moving the computer mouse. Directly following this, they were required to report the amount of points they were willing to bet that the answer they gave was the correct one. Coherence of the stimulus was experimentally manipulated and participants were required to complete 11 experimental blocks, each containing 48 trials of varying dot coherence. Mouse trajectory information was not predictive of post-decision certainty but was strongly related to decision accuracy. The findings were in line with a view of confidence as an evaluation of evidence which continues to accumulate after a decision.

Typically in our day to day lives, as we make a judgement or decision, it is followed by a subjective sense of certainty that the right decision was made. This feeling of certainty in our choices is commonly termed decision confidence. In the context of perceptual decision making, two current accounts suggest different mechanisms behind formation of confidence. Van den Berg et al. (2016) argue that confidence arises from the same evidence accumulation mechanism that underlies the formation of the original decision. In contrast, Murphy et al. (2015) propose that confidence is a product of a higher-order meta-cognitive process evaluating evidence beyond the initial decision. Most recently, Fleming and Daw (2017) proposed a model of second-order decision confidence, which generalizes and unifies these two approaches.

The current study investigated whether characteristics of participant's hand movements can act as an objective measure of on-line decision confidence. Comparing this measure to both performance accuracy and subjective confidence reports provides further insight into the process underlying retrospective confidence. Participants were required to make a perceptual discrimination task under differing levels of certainty and subsequently supply a measure of post hoc confidence. While previous research has often looked exclusively at associations between performance on a task and subsequent confidence judgements, the current study looks at associations between performance accuracy, confidence reports, and response trajectories. In this way, this research aims to provide another layer of evidence to the debates on the nature of decision confidence.

Methods

Undergraduate psychology students ($N = 28$, 4 male, 24 female, mean age 19.6 years) completed an experimental session in exchange for course credit. All study procedures

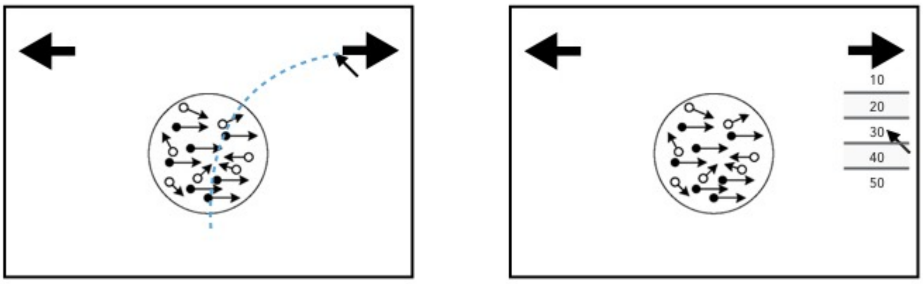


Fig. 1. Experimental setup: response screen during initial decision (left panel) and gamble selection (right panel)

employed were approved by the Research Ethics Committee at the National University of Ireland, Galway.

Participants made a series of 528 perceptual decisions, each followed by a “gamble”. On each trial, participants were asked to judge the prevalent direction of a random dot kinematogram (RDK). The dot generation algorithm employed in this study was based on that of Shadlen and Newsome (2001) and programmed using PsychoPy (Peirce, 2007). The participants indicated their choice by moving the mouse cursor from the starting position at the bottom centre and clicking on one of the response locations in the top corners of the screen (Fig. 1). The RDK stimulus was present on the screen until one of the response locations was selected. Mouse coordinates during the response were recorded at 60Hz.

Motion coherence (the probability of any particular dot being displaced in the stimulus direction) constituted one independent variable and was manipulated within participants. The experimental session required participants to complete 11 trial blocks, each consisting of 48 trials. There were four coherence levels presented within each block (0.032, 0.064, 0.128, 0.256). All experimental blocks contained 12 trials of each coherence level, randomly shuffled, with stimulus direction (left or right) randomly determined for each trial.

On choosing a direction, participants were required to gamble 10 to 50 points on their answer using a drop-down menu. If the correct direction was chosen, a participant gained the chosen points, and if they chose the incorrect direction, the same number of points was deducted from their accumulated score. Participants were told at the beginning of the experiment that the aim was to earn as many points as possible by the end of the session. Upon making a number selection, a feedback stimulus appeared on the screen informing subjects whether they had answered correctly and the number of points won or lost.

Results and Discussion

The present analysis excludes the participants who could not perform the task accurately enough. Specifically, we excluded 12 participants who had accuracy below 75% at the coherence level of 0.256 and/or accuracy below 65% at the 0.128 coherence level.

The rest of the participants ($N = 16$), as expected, were more likely to correctly discriminate the direction of the RDK stimulus with increasing coherence (Fig. 2). In approximately 50% of the trials, the mouse trajectories indicated that the participants changed their preference during the course of the trial. Such trials were labelled as

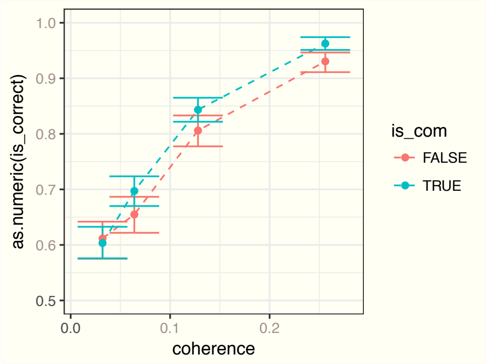


Fig. 2. Psychometric functions (averaged across 16 participants) for change-of-mind and non-change-of-mind trials.

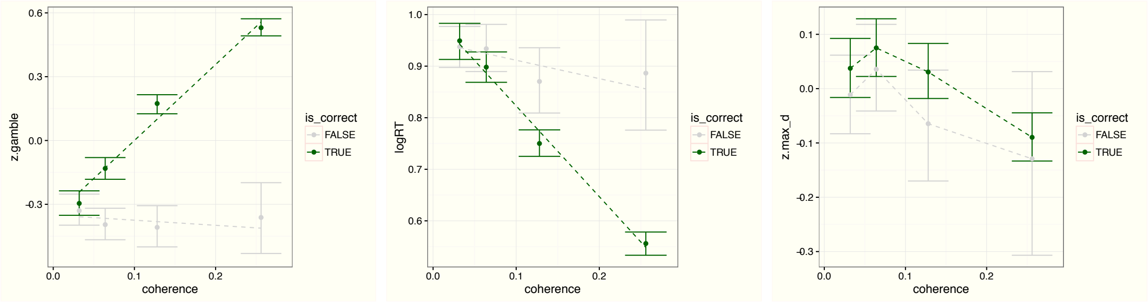


Fig. 3. Gamble value (z-scored), response time (log-scaled), and mouse trajectory curvature (z-scored) as a function of coherence for correct and incorrect trials. Correct responses are depicted in green and incorrect responses in grey. Whiskers denote bootstrapped 95% confidence intervals.

changes-of-mind. In accordance with the previous studies (e.g., Resulaj et al., 2009), changes-of-mind improved accuracy (Fig. 2). This might reflect the fact that after initial decision, additional evidence was continuously available to the participants until they clicked on one of the response locations.

The amount of points gambled after each decision increased with coherence for correct trials (Fig. 3, left panel). For incorrect trials, the amount gambled remained consistently small across coherence levels, which indicates that the participants could reliably detect their erroneous responses post-decision when the stimulus coherence was high (0.128 or 0.256). Together, these patterns suggest that gambled amount reflects subjective post-decision confidence of decision makers.

Previously in the mouse-tracking literature on value-based decision-making it has been suggested that mouse trajectories are linked to relative subjective value of the available options (e.g., McKinstry et al., 2008; Dshemuchadse et al., 2013; O'Hora et al., 2016). Here we hypothesize that in perceptual decision making, mouse trajectories may provide a measure of confidence within the response. To this end, we analyse response time and trajectory curvature as a function of RDK coherence.

As expected, response time decreased with coherence for correct responses (Fig. 3, centre panel). Error response time tended to remain high for all coherence levels. This is in line with consistently high values of gamble value, and thus reinforces the view of response times being related to decision confidence.

To measure trajectory curvature, we calculated maximum deviation (max-d) of each trajectory from the shortest trajectory towards the corresponding response area. In correct trials, max-d for correct trials exhibited non-monotonic relationship with coherence (Fig. 3, right panel). Although one might expect trajectory curvature to decrease as coherence increased (that is, as decisions became easier), mean max-d initially increased as coherence increased from 0.032 to 0.064. When coherence increased further to 0.128 and then 0.256, mean max-d decreased, in line with the expected pattern. As higher values of max-d in the present paradigm indicate greater rate of changes-of-mind, this finding is consistent with the post-decision evidence accumulation account of changes-of-mind (Resulaj et al., 2009).

Overall, our results suggest that changes of response direction during motor execution of a decision are informed by late-coming signal rather than by noise in the stimulus. Moreover, decision confidence as reflected in post-decisional wagering is related, but not equivalent to curvature of the response trajectories. Further investigations will shed light on the nature of relationship between within- and post-decision confidence.

Acknowledgements

This research was sponsored in part by Government of Ireland Postdoctoral Fellowship GOIPD/2015/481.

References

Dshemuchadse, M., Scherbaum, S. and Goschke, T. (2013), ‘How decisions emerge: Action dynamics in intertemporal decision making.’, *Journal of Experimental Psychology: General* **142**(1), 93.

Fleming, S. M. and Daw, N. D. (2017), ‘Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation.’, *Psychological Review* **124**(1), 91.

McKinstry, C., Dale, R. and Spivey, M. J. (2008), ‘Action dynamics reveal parallel competition in decision making’, *Psychological Science* **19**(1), 22–24.

Murphy, P. R., Robertson, I. H., Harty, S. and O’Connell, R. G. (2015), ‘Neural evidence accumulation persists after choice to inform metacognitive judgments’, *Elife* **4**, e11946.

O’Hora, D., Carey, R., Kervick, A., Crowley, D. and Dabrowski, M. (2016), ‘Decisions in motion: Decision dynamics during intertemporal choice reflect subjective evaluation of delayed rewards’, *Scientific Reports* **6**, 20740.

Peirce, J. W. (2007), ‘PsychoPy—psychophysics software in Python’, *Journal of Neuroscience Methods* **162**(1), 8–13.

Resulaj, A., Kiani, R., Wolpert, D. M. and Shadlen, M. N. (2009), ‘Changes of mind in decision-making’, *Nature* **461**(7261), 263–266.

Shadlen, M. N. and Newsome, W. T. (2001), ‘Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey’, *Journal of Neurophysiology* **86**(4), 1916–1936.

Van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N. and Wolpert, D. M. (2016), ‘A common mechanism underlies changes of mind about decisions and confidence’, *Elife* **5**, e12192.