Short Report

PERCEPTION

Visual Search: You Are Who You Are (+ A Learning Curve)

Perception
2017, 0(0) 1–8

DOI: 10.1177/0301006617721091
journals.sagepub.com/home/pec
SAGE

Justin M. Ericson
The George Washington University, Washington, DC, USA;
Duke University, Durham, NC, USA

Dwight J. Kravitz and Stephen R. Mitroff
The George Washington University, Washington, DC, USA

Abstract

Not everyone is equally well suited for every endeavor—individuals differ in their strengths and weaknesses, which makes some people better at performing some tasks than others. As such, it might be possible to predict individuals' peak competence (i.e., ultimate level of success) on a given task based on their early performance in that task. The current study leveraged "big data" from the mobile game, *Airport Scanner* (Kedlin Company), to assess the possibility of predicting individuals' ultimate visual search competency using the minimum possible unit of data: response time on a single visual search trial. Those who started out poorly were likely to stay relatively poor and those who started out strong were likely to remain top performers. This effect was apparent at the level of a single trial (in fact, the first trial), making it possible to use raw response time to predict later levels of success.

Keywords

individual differences, perceptual learning, search or exploration, visual search

Introduction

Individuals differ in their relative strengths and weaknesses, with certain individuals performing better on specific tasks in comparison to others. Such variability is natural and commonplace, but it is nevertheless interesting to explore why some individuals are "just better" than others at a task. One key issue for understanding why some individuals excel at a task while others struggle is whether or not task competency is stable—do those who reach high levels of performance also start out relatively stronger? This issue represents the primary focus of the current study; examining how individuals' peak competence (i.e., their eventual performance threshold) is related to their initial performance on a task. The ability to predict peak competence could have broad theoretical and practical

Corresponding author:
Justin M. Ericson, George Washington University, 2125 G St., NW, Washington, DC 20052-0086, USA.
Email: justin.m.ericson@gmail.com

implications, especially if peak competence can be predicted from a small sample of early performance. Theoretically, it could reveal the degree to which cognitive performance is driven by participants' preexisting abilities, suggesting that the relative performance of some behaviors across individuals may not be easily shaped or manipulated. Moreover, this presents an intriguing methodological and academic question—what is the minimal amount of information needed to reliably predict later success? Practically, it could be financially and logistically powerful to take a small sample of data and have some reliable evidence about which individuals are most likely to eventually become star athletes, Navy Seals, pilots, and so forth. Personnel selection (e.g., Ryan & Ployhart, 2014; Sackett & Lievens, 2008) is an exciting interface between basic and applied research, and it is intriguing to consider just how little data could be used to predict employment success.

The current study explored whether it is possible to identify, from a minimal amount of early performance, which individuals are consistently better at *visual search*—the ability to detect targets among distractors. Search requires critical aspects of cognition (e.g., attention, perception, and memory; for recent reviews, see Eckstein, 2011; Nakayama & Martini, 2011) and underlies many real-world tasks (e.g., airport security, radiology, and lifeguarding). Thus, identifying individual differences in performance, and predicting eventual competence, can ultimately inform both academic theories and improve real-world applications. Generally speaking, performance on a task should improve with experience (and it does; Stafford & Dewar, 2014), but a key question is whether individuals' *relative* competence is maintained over time. That is, do those who start relatively poor, mediocre, and excellent remain relatively poor, mediocre, and excellent throughout? While it seems a given that individuals will start a task at different levels of aptitude (i.e., some will be better on their first try), what is not known is how stable those differences are from initial to later performance.

The methods for predicting peak competence vary in (a) the amount of early performance data needed to make the prediction, (b) the complexity and fidelity of the measure of peak performance, and (c) the complexity and assumptions which underlie the model that is used to relate early to peak performance. As an academic exercise to determine the minimal amount of information needed for prediction, the current study implemented a minimal and assumption-free version of all three of these factors. Specifically, response time taken from a single trial was used to predict an individual's later, peak performance. That is, the minimal unit of measurement from performance in a visual search task (how fast a participant responds on a single trial) was used to predict the participant's level of competency that was later achieved in the search task. This minimalist approach serves two goals: First, it provides an extremely aggressive test of the hypothesis that early performance can predict peak competence; Second, it provides a lower bound of potential informativeness—any relationship derived from this minimal level of analysis would represent the base level of predictability; as additional processing of the data would likely reveal even stronger relationships. Given this minimalist approach, this study serves as a proof of concept about the ability to predict later relative competence from early performance—in practice (say, for example, in predicting who might be an appropriate professional visual searcher for medical image perception or airport security), it would make sense to include more than just a single trial as a basis for the prediction.

Methods

Data were analyzed from a visual search dataset derived from the mobile technology game *Airport Scanner* (Kedlin Company; see Mitroff et al., 2015). *Airport Scanner* is a publicly available game on iOS and Android devices, with players assuming the role of airport x-ray baggage screeners tasked with searching through simulated bags (constituting trials) for prohibited items. As of November 2016, there have been over 11 million installs of the app providing a dataset that contains over 2.8 billion individual trials (for more detail on the nature of *Airport Scanner* and it's use for psychological research, see Biggs, Adamo, & Mitroff, 2014; Biggs, Adamo, Dowd, & Mitroff, 2015; Mitroff & Biggs, 2014; Mitroff et al., 2015). The game begins simple with only two to-be-searched-for target items but becomes progressively more complex as the players advance, reaching up to 225 different targets. Players begin with a "rank" of *trainee*, and then advance in rank (awarded in the order *Trainee*-*Operator*-*Pro*-*Expert*-*Elite*) as they successfully pass harder and harder levels. Rank was used as the primary outcome measure in the current analyses, with participants categorized as those who reached their ultimate rank of *pro*, *expert*, or *elite*.

A subset of the entire dataset was used for the current purposes. From this subset, 80,760 individual players met the inclusion criteria of having reached the *Pro* rank and had completed at least 180 target-present and 180 target-absent trials following the introductory training and practice levels. The 180 trial limit was selected as this provided a number of trials needed for a player to be well within the *Pro* rank for inclusion. This inclusion criteria eliminated players who only played for a few trials, or who did not make it past the early game ranks, and ensured at least somewhat extended play for all included players. All performance data were gathered from the individuals who began playing *Airport Scanner* between March 15, 2013 and October 1, 2014, recording their first trial attempted, to their last trial within that period. Their ultimate rank (*Pro*, *Expert*, *Elite*) was determined by their final status within this date range for using the app. Additional filters removed trials with active in-game upgrades that aid search performance as well as any trials that had response times below 250 ms.

The competence measure used in the current study was simply the highest rank achieved in the game. This is a broad delineation and is open to a wide variety of potential noise (e.g., players can stop playing at will for a variety of reasons). For the current research purposes, the grouping of players was specifically chosen to match a simple predictive measure (single trial response time) with a simple outcome measure (rank achieved). This truly tests how little information is required to reliably predict later success.

Results

The primary analysis assessed trial-by-trial response time for those trials in which the player had accurately identified the bag as either containing a target (target-present response time: time to tap target) or not containing a target (target-absent response time: trial duration). The goal of this analysis was to use the smallest possible unit of performance—a single trial—to predict later outcomes. Mean response times for target present (Figure 1(a)) and target absent (Figure 1(b)) trials revealed group-level differences. Specifically, those who ultimately achieved the highest rank of *Elite* were already on average faster from the first trial and their advantage over the other groups held for every trial assessed (i.e., across all 180 trials for both target-present and target-absent analyses). Likewise, the players who ultimately achieved *Expert* rank were faster on every trial compared with those who achieved only the *Pro* rank.

To establish whether a single trial of early performance was sufficient to reliably predict a player's ultimate rank (i.e., their peak level of achievement), the first step was to sort performance for each trial by the eventual rank of the player and whether the trial contained a target or not (Figure 1). The resulting groups of response times

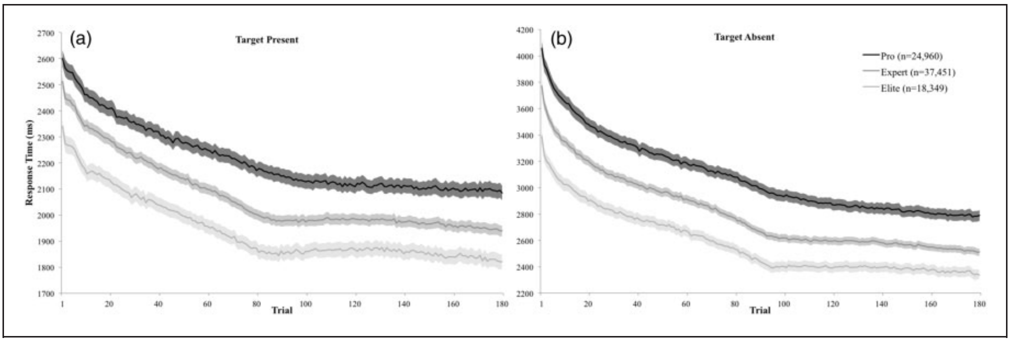

Figure 1. Mean RT by trial for target present (a) and target absent (b) trials, sorted by the maximum rank achieved. Shaded areas represent 99.9999999% confidence intervals (six between-subjects standard errors of the mean).

(target-present or target-absent) were then each subjected to a two-way mixed ANOVA with trial number (180) as a repeated measure and eventual rank (*Pro*, *Expert*, *Elite*) as a between-subjects factor. The target-present ANOVA revealed significant main effects of trial number, $F(179, 1445496) = 4{,}742.47$, $p < .0001$, and, critically, eventual rank, $F(2, 80754) = 5{,}623.29$, $p < .0001$. A significant interaction between rank and trial number was observed, $F(358, 1445496) = 19.80$, $p < .0001$, suggesting that some rank groups showed greater changes in performance across trials. To verify that performance across rank was significantly different for each trial, a series of pairwise *t*-tests between every pairing of ranks was conducted for each trial. Even following a Bonferroni correction for all 1,080 comparisons ($\alpha = .000046$), significant differences were observed between all ranks for all trials (all p's $< 2.37 \times 10^{-35}$). The results of the target-absent ANOVA revealed similar effects, demonstrating that performance was similar across trial types. There were again significant main effects for trial number, $F(179, 1445496) = 2{,}831.19$, $p < .0001$, and for eventual rank, $F(2, 80754) = 3769.01$, $p < .0001$. Additionally, there was also a significant interaction, $F(358, 1445496) = 7.128$, $p < .0001$.

The results of these abovementioned analyses highlight that it is possible to observe highly significant average differences between the ranks, from even the very first trial, $t(57511) = 16.45$, $p < .0001$. However, the analyses reveal very little about the performance of any given individual. Upon first glance, the miniscule size of the standard errors relative to the difference between the ranks, and the accompanying *p*-values, might suggest that predicting the rank of an individual participant would be trivial. However, these statistics only measure the likelihood that the *average* performance taken from each rank is identical, not whether any individual player can be accurately associated with their eventual rank. The next analysis demonstrated that even at the level of individual participants, it was still possible to predict rank based on single trials using a simple model. Due to the inherent variability of target-absent trials, with exhaustive search having a wide range of end points, only target-present trials were used for the subsequent model analysis.

Predictive Analysis

To determine if early performance predicted peak visual search competence (as defined by the highest rank achieved by the player), data from 15,000 players from each rank were taken and divided into three independent sets of 5,000 each. In an iterative leave-one-out

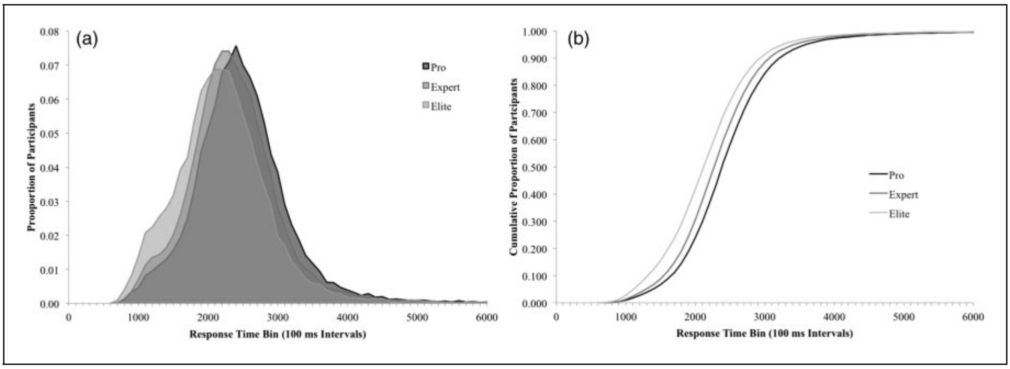

Figure 2. Target present RT distributions (binned in 100 ms intervals) across participants for the 73rd target present trial (the trial that showed the greatest separation between ranks). (a) Proportion of participants within a RT bin grouped by rank (probability density function). (b) Cumulative density function of the same distribution.

procedure, two of these sets were combined to generate a model set (10,000 players from each rank). The remaining 5,000 players were held out so that they could be used for testing the prediction. Then, a criterion was derived from each model set that maximally discriminated between each possible pairing of ranks, as described later. The efficacy of this criterion was tested against the 10,000 relevant participants (5,000 per rank).

Player response times from the model set for each trial were binned in 100 ms intervals to create histograms (e.g., Trial 73 in Figure 2). As can be seen from the example trial (Figure 2), the distributions demonstrated some overlap, but clear differences between each of the three ranks (*Pro*, *Expert*, and *Elite*).

Using the model distributions from each trial (e.g., Trial 73, Figure 2), an optimal response time criterion was determined that maximally discriminated between each pairing of ranks (e.g., best separation between *Expert* and *Elite* players). For each trial and for each comparison between ranks, the remaining 5,000 players per rank from the test set were evaluated against the response time criterion from the model set. That is, each player in the test set was evaluated using the criterion from the model set to estimate which rank they were most likely to fall into based on their response time. This was done separately for each trial. The probability of accurate prediction was calculated as the average proportion of the test set players (5,000 per rank) that fell on the correct side of the criterion established in the model set (Figure 3). Following this procedure, a series of one-sample *t*-tests were performed to examine whether the probability of each comparison was different from chance (50%). Analysis revealed that even with a Bonferroni adjusted alpha level for 540 comparisons ($\alpha = .00009$), performance was above chance for all comparisons at all trials (all p's < .0000000001), including the very first trial.

The result of this predictive analysis suggests that it is possible to determine eventual competence from a single trial of data, even when it is the first trial. Given the correlational nature of these analyses, there is no means or reason to claim that the early performance caused the later success, but rather it is meaningful that early performance, even from a single trial, predicted later performance. With predictions from single trials being at least 55% correct, it may be possible to gain even greater predictive power by investigating performance from more than just a single trial in isolation. The abovementioned analysis represents the bare minimum of input (i.e., response time on a single trial), and obviously it

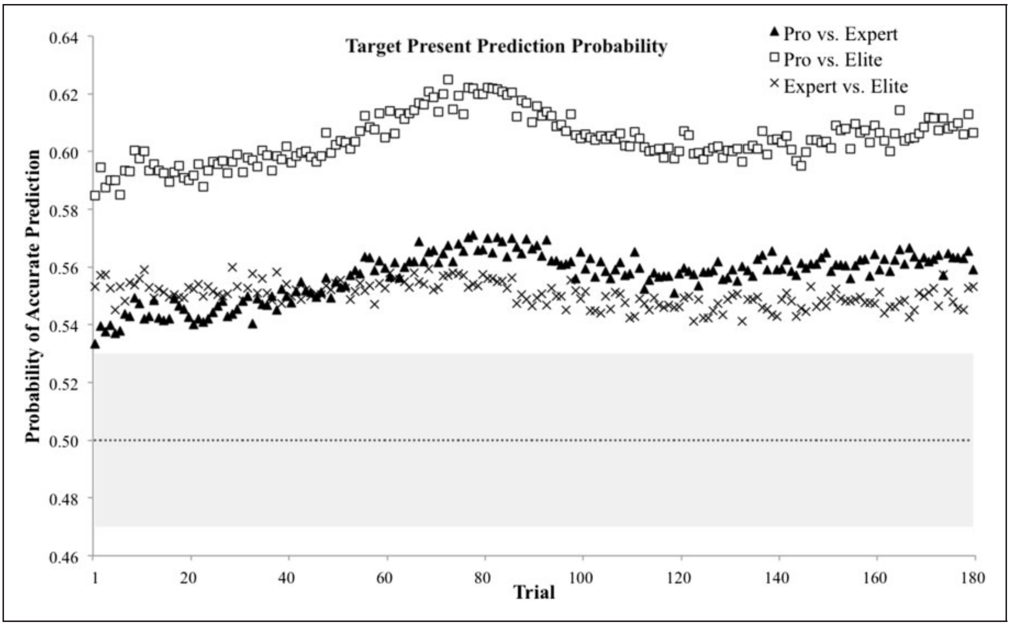

Figure 3. Probability for accurate prediction between eventual ranks evaluated at each trial independently for target present trials. The dashed line indicates chance performance and the gray area around it is the 99.9999999% confidence interval (six between-subjects standard errors of the mean).

is possible to incorporate more data and more complexity (e.g., accuracy and response time, sensitivity, bias, data across multiple trials).

To demonstrate the ultimate power of this line of inquiry, later we take an additional simple step—averaging performance as it accumulates across trials. The prediction is that taking the accumulative average response times from trial-to-trial would not only bolster the predictability of an individual's eventual rank but would also identify the minimum number of trials required to reach the highest level of prediction. The procedure was identical to that of the previous analysis except that rather than take the individual response times from each trial, response times were averaged across all 180 trials and subjected to the same iterative procedure described earlier.

Like the single-trial analysis, the response time distributions from this accumulative model were used to establish a criterion to discriminate between each pairing of ranks. The test set was then evaluated against the model criterion in the same manner as the previous analysis. The probability of accurate prediction was calculated as the proportion of the test set players that fell on the correct side of the criterion established in the model set (Figure 4). Using the average response time resulted in a maximum accuracy of 71%. These findings demonstrated how prediction probability could be easily improved by taking the average response times across trials. Moreover, this simple analysis—averaging across trials—highlights the potential of the current efforts, as more and more performance data can be included to increase the predictive power of the models.

General Discussion

These results demonstrate clear differences in early performance between individuals sorted by a simplistic rule—their ultimate rank achieved in *Airport Scanner*. All players received the

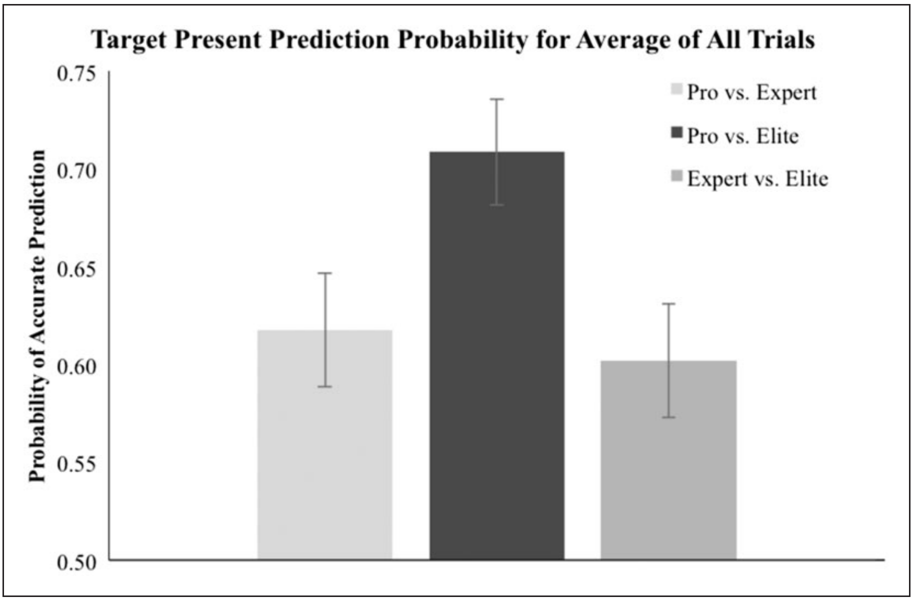

Figure 4. Probability for accurate prediction between eventual ranks evaluated for the average response time across the first 180 target present trials. Error bars represent 99.9999999% confidence intervals (six between-subjects standard errors of the mean).

same amount of instruction and practice; however, those who obtained a higher rank in the game, on average, began with better performance from even the very first trial and continued to maintain a more efficient search. Furthermore, players' performance on single trials (even the first trial) was predictive of their eventual rank based on a model derived from independent players consisting of a single-fitted parameter—a criterion response time.

The current study investigated a sample of data utilizing only response time as a measurement, which resulted in an accurate prediction of later competence in a visual search task. Future work can combine data (e.g., accuracy and response time across trials) to produce more predictive markers of eventual achievement. Other individual difference measures could also be used to accurately predict outcomes or be used to identify those individuals who likely show stronger early performance. Additionally, the current study utilized data taken from a single application (*Airport Scanner*), and it is exciting to consider how combining data from across multiple applications could provide a wealth of potential information that can serve the cognitive sciences. Assessing data originating from multiple sources or attributes taken from daily performance could easily result in predictions that exceed the 71% accuracy reported here between our *pro* and *elite* players. The primary goal of the current study was to explore the theoretical question of whether a single trial can predict peak visual search competency. Future work can build on this academic finding to address more practically relevant questions, such as just how strong of a prediction can be obtained when combining multiple sources of data.

In conclusion, early search performance can serve as a reliable predictor of eventual peak search competence. The current findings suggest that it might be possible to eventually use even small amounts of data to guide personnel selection (e.g., inform airport security personnel hiring) and training, assessment, or intervention strategies. These results suggest that relative ability may be a stable aspect of performance, as the differences between the participant groups remained constant across trials. However, performance still became much more efficient across trials for all groups (Figure 1), suggesting an avenue remains for training, interventions, and different rewards to make the learning curve more efficient. Interestingly, the stability of the absolute difference in performance between the groups causes the proportional advantage of the top performers to grow. Thus, in large organizations, better personnel selection may have proportionally greater impact on overall average efficiency following training. Many additional questions arise from the current study, but perhaps the most intriguing is to examine what causes certain individuals to deviate from the predictive rule discussed here and outperform or underperform their initial relative performance, allowing for the design of better interventions.

Acknowledgements

The authors thank Ben Sharpe, Thomas Liljetoft, and Kedlin Company for access to the *Airport Scanner* data. The authors also thank Christina Gancayco and members of the Duke and GW Visual Cognition labs for their assistance.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially funded by Army Research Office awards #W911NF-14-1-0361 and W911NF-16-1-0274.

References

Biggs, A. T., Adamo, S. H., & Mitroff, S. R. (2014). Rare, but obviously there: Effects of target frequency and salience on visual search accuracy. *Acta Psychologica*, *152*, 158–165.

Biggs, A. T., Adamo, S. H., Dowd, E. W., & Mitroff, S. R. (2015). Examining perceptual and conceptual set biases in multiple-target visual search. *Attention, Perception, & Psychophysics*, *77*, 844–855.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, *11*, 1–36.

Mitroff, S. R., & Biggs, A. T. (2014). The Ultra-Rare-Item effect: Visual search for exceedingly rare items is highly susceptible to error. *Psychological Science*, *25*, 284–289.

Mitroff, S. R., Biggs, A. T., Adamo, S. H., Dowd, E. W., Winkle, J., & Clark, K. (2015). What can 1 billion trials tell us about visual search? *Journal of Experimental Psychology: Human Perception & Performance*, *41*, 1–5.

Nakayama, K., & Martini, P. (2011). Situating visual search. *Vision Research*, *51*, 1526–1537.

Ryan, A. M., & Ployhart, R. E. (2014). A century of selection. *Annual Reviews of Psychology*, *65*, 693–717.

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annual Reviews of Psychology*, *59*, 419–450.

Stafford, T., & Dewar, M. (2014). Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological Science*, *25*, 511–518.