Neth, H., Sims, C.R., Veksler, V.D. & Gray, W.D. (2004). You can't play Straight TRACS and win: Memory updates in a dynamic task environment. In K.D. Forbus, D. Gentner & T. Regier (Eds.), Proceedings of the Twenty-Sixth Annual Meeting of the Cognitive Science Society (pp. 1017-1022). Hillsdale, NJ: Lawrence Erlbaum.

You Can't Play Straight TRACS and Win: Memory Updates in a Dynamic Task Environment

Hansjörg Neth (nethh@rpi.edu) **Chris R. Sims** (simsc@rpi.edu) **Vladislav D. Veksler** (vekslv@rpi.edu) **Wayne D. Gray** (grayw@rpi.edu)
Cognitive Science Department
Rensselaer Polytechnic Institute

Abstract

To investigate people's ability to update memory in a dynamic task environment we use the experimental card game TRACS™ (Burns, 2001). In many card games card counting is a component of optimal performance. However, for TRACS, Burns (2002a) reported that players exhibited a baseline bias: rather than basing their choices on the actual number of cards remaining in the deck, they chose cards based on the initial composition of the deck. Both a task analysis and computer simulation show that a perfectly executed memory update strategy has minimal value in the original game, suggesting that a baseline strategy is a rational adaptation to the demands of the original game. We then redesign the game to maximize the difference in performance between baseline and update strategies. An empirical study with the new game shows that players perform much better than could be achieved by a baseline strategy. Hence, we conclude that people will adopt a memory update strategy when the benefits outweigh the costs.

Introduction

Optimal performance in dynamic environments requires that we base our decisions on the current state of the world, not on past states. Radar operators must act on the basis of continuously changing variables such as plane altitude and heading. Drivers constantly need to monitor the current speed limit, posted road signs and the traffic behind and in front of them. Failure to mentally update these types of information can lead to dangerous decisions and catastrophic behavior. Even our chances to win at card games like Blackjack or Bridge are closely tied to our ability to count cards and update memory.

Previous research suggests that human ability to monitor and adjust to change is limited and dependent on various factors. Yntema (1963) found that people are better at tracking a small number of variables with a large range of values each, than a large number of variables with a small set of possible values each. In addition, reducing the frequency of update can improve performance. Other manipulations, such as increased predictability of a sequence, provide little or no advantage in remembering the current state of the environment. Venturino (1997) distinguished the memory capacity for static information from that for dynamically changing information and showed that the latter is highly limited, particularly when the to-be-remembered attributes are similar. Hess, Detweiler and Ellis (1999) added that update performance is improved when spatial invariants constrain where different data values are presented on a visual display.

In general, human rational behavior is constrained by the structure of task environments and the computational capabilities of the actor (Simon, 1990). To capture functional relationships of complex tasks while abstracting away from domain specific details we advocate the use of synthetic task environments, or microworlds (Gray, 2002). If the properties of the synthetic task environments are known and manipulable, the scope and limits of human rationality can be assessed. Moreover, the effects of environmental changes are tractable.

Table 1: Baseline distribution of cards in the deck. The back of every card shows only its shape, whereas the front shows both its shape and color.

Shape:	▲	●	■	▲	●	■
Color:	red	red	red	blue	blue	blue
Initial deck:	6	4	2	2	4	6

Straight TRACS

TRACS™ is a 'Tool for Research on Adaptive Cognitive Strategies,' designed and developed by Kevin Burns (2001, 2004). Being both entertaining card game and experimental research tool, TRACS provides a microworld which promises to bridge the gap between mathematical rigor and real-world relevance. We will limit our discussion to *Straight* TRACS, which is the simplest version of an entire family of games.[1]

TRACS is played with a deck of 24 cards. The back of each card shows one of three shapes—circle, triangle, or square—filled in with black. The front of each card shows both its shape, and one of two colors (red or blue). Table 1 shows the initial deck distribution for each of the six possible card types. This baseline information is always available to the player. As hands are played the number of cards remaining in the deck decreases, and the odds for each shape change accordingly.

At the start of a game, three cards are dealt in a row. The middle card is dealt face up (showing both its shape and color), while the left and right cards are dealt face down, showing their shape not their color. The task for the player is to choose the card, either left or right, most likely to match the *color* of the middle card. The chosen card is then turned over, revealing its color. If the chosen card matches the color of the middle card, a *hit* is credited to the player's score. A mismatch is scored as a *miss*. The two face up cards (the middle and the chosen card) are then removed from the game. On

[1] Online versions are available at www.tracsgame.com.

the next turn, the unchosen card is flipped over and becomes the new middle card, and two new cards are dealt face down to the left and right. A game lasts 11 turns, at which point there are not enough cards in the deck to deal another hand. A player's objective in TRACS is to maximize the number of hits.

As a probe of the player's assessment of odds at each turn, Burns (2002a, 2002b) added a confidence meter to the task. On each turn, players were presented with a red to blue color gradient for each of the two face-down cards. Prior to choosing a card, the participants used the gradient to indicate the likelihood of each candidate card to be red or blue. In another condition, Burns used a scale of nine buttons rather than a continuous spectrum. For consistency reasons all gradient estimates were rounded to the nearest button, corresponding to the nearest 12.5%.

Burns (2002a) characterized players' likelihood estimates as exhibiting a *baseline bias*; i.e., their judgments of odds deviate systematically from the actual odds in the direction of the initial card distribution. There are six types of color–shape combinations. Burns (2002b) reports that players could only monitor 2–4 types of cards with reasonable reliability. He concludes that the dual tasks of concurrently counting and normalizing numbers 'are naturally hard' and that continuously updating odds exceeded the cognitive capacity of the 'unaided mind' (Burns, 2002a, p. 159).

In the following sections, we will challenge this claim both theoretically and empirically. To preview our conclusions, we find that subtle constraints in the task environment can have profound effects on the strategy adopted by participants. The reported baseline bias is revealed as both rational and adaptive when considered in light of a cost-benefit analysis of the environment. We then demonstrate that players will adopt a more effortful memory strategy if the cost-benefit structure of the environment rewards this.

Tracking TRACS

Given the original finding that players find it challenging to succeed at TRACS, a natural starting point for our investigation is a task analysis. What specifically makes this game so difficult to play?

Task Analysis

In describing TRACS as a game of 'confidence and consequence' Burns (2001, 2002b) distinguishes two subtasks of diagnosis and decision. On each turn, a player first provides an odds judgment for each face-down card and then chooses one on the basis of these estimates.

Extending Burns' analysis, we suggest that each turn involves a minimum of *three* distinct cognitive tasks: a memory retrieval task, an odds conversion task, and a decision task (see Figure 1). The first subtask on each turn consists in remembering how many cards of each candidate shape and color remain in the deck. As the initial card distribution is provided in terms of frequencies and players encounter card instances through a process of natural sampling, we assume that this retrieval is framed in terms of natural frequencies. Secondly, the retrieved frequencies need to be converted into odds, which is a non-trivial process involving Bayes' rule for natural frequencies (Gigerenzer, 2000). For example, to determine the likelihood of a red triangle, a player has to divide the number of red triangles currently left in the deck by the sum of red and blue triangles left in the deck. As people are notoriously bad at dealing with probability information (see Gigerenzer, 2000, and Koehler, 1996, for reviews) it is conceivable that this translation process incurs a loss of accuracy. If so, merely asking for likelihood estimates confounds memory updates with probability judgments and may underestimate players' true memory capacity. As a third subtask, a player needs to integrate all estimates and decide which candidate card is more likely to score a hit on the current trial.

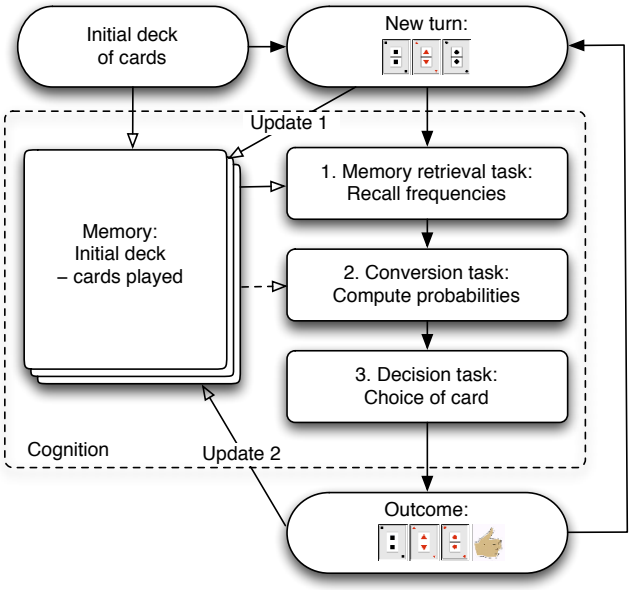

Figure 1: Subtasks and memory updates required on each turn of Straight TRACS.

In addition to these three subtasks, each turn requires two distinct updates of memory. The first update is necessary as soon as the middle card is revealed. If the middle card happens to be a red triangle, the player needs to realize that there now is one less red triangle left in the deck. The second update ought to occur at the end of a turn when the chosen card is revealed. This second update is critical, as at this point in the game, players may be distracted by focusing exclusively on the correctness of their choice and ignoring the additional information revealed.

This task analysis reveals both the complexity and simplicity of TRACS. On one hand, multiple subtasks and memory update requirements make the game quite challenging. Even if frequency information on card types was readily available, the conversions into probabilities, comparisons between odds, and selection of cards introduce potential sources of error. On the other hand, remembering and updating a list of six numbers (representing the current frequency of each card type) does not in itself seem beyond the capacity of human memory.

The Impact of Memory

At first glance, it seems that TRACS is a 'memory game' (Burns, 2001, 2002a) in which players can succeed only by remembering which cards have left the deck. However, our

experience playing TRACS casts some doubts on the importance of memory. Due to the random card selection process a typical game contains many knowledge-indeterminate turns. For example, whenever both face-down cards show the same shape, a player has no choice but to guess. Likewise, both face-down cards frequently have the same color, so that the player scores a hit or miss regardless of knowledge or choice. Even when the cards differ in shape, color, and odds, it is possible that selecting the card with higher actual odds results in a miss, whereas choosing the 'wrong' card scores a hit.

These concerns raise questions about whether memory really matters. To what extent can poor performance be blamed on failures of memory? Would better memory improve performance? The non-deterministic nature of the game makes it hard to answer these questions analytically; thus, we implemented the game as a computer simulation.

Simulation As Allen Newell and Herbert Simon famously stated, "Just as a scissors cannot cut paper without two blades, a theory of thinking and problem solving cannot predict behavior unless it encompasses both an analysis of the structure of task environments and an analysis of the limits of rational adaptation to task requirements." (1972, p. 55). In this spirit, we created a simulation in MATLAB™ in which 'pure' cognitive strategies could be formalized and implemented. By running these artificial agents for thousands of trials, we were able to determine precise performance levels, despite the dynamic and nondeterministic aspects of the game.

We compared four cognitive agents that differed in their memory resources and strategies, but did not make any errors in odds translation or judgment. A *baseline* agent has perfect knowledge of the initial deck distribution, but is amnesic with regards to the cards played during a game. In contrast, the *update* agent enjoys perfect memory of every hand played, and bases all choices on the actual odds at any given moment.

Two additional agents bracket the performance of baseline and update agents: *random* agent has neither memory nor knowledge of the initial distribution, and hence is forced to blindly guess at every turn. On the other end of the scale, *omniscient* agent effectively enjoys X-ray vision and can observe the colors of both candidate cards, allowing for optimal card selections without the need for memory or odds estimates.

The mean score for the random agent across 10,000 simulated games was 5.24 (out of 11 possible) hits per game. To our surprise, baseline and update agents performed about the same, scoring 6.57 and 6.79, respectively. Thus, the average performance difference between the baseline and update agents was roughly two tenths of one point per game. Further, both strategies achieved only marginally better scores than the random strategy.

Figure 2 shows the mean percentage of hits per turn for each agent. It is obvious that the performance of baseline and update agents are very similar, except for an increasing benefit of update strategy late in a game. The entire range between random and omniscient performance scores is only 25%, which is essentially due to 25% of all turns not allowing for a hit.

While an optimal update agent acts to maximize performance regardless of the effort involved, humans have limited cognitive resources and are required to negotiate cost–benefit tradeoffs (Anderson, 1990; Simon, 1990, 1992). Given these

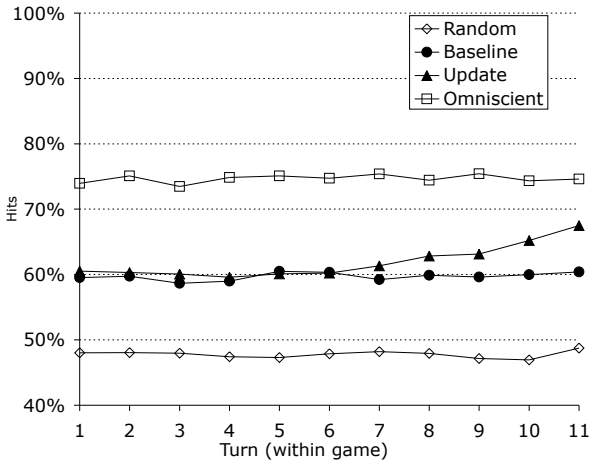

Figure 2: Simulation results for four artificial agents playing 10,000 games of original TRACS.

constraints and the minimal benefits of an update strategy, participants might well have adopted a baseline strategy for good reasons. Thus, our analysis suggests a re-interpretation of Burns' original findings: In Straight TRACS, memory update yields no performance benefit over adopting a much easier baseline strategy. Hence, adopting the baseline strategy is both adaptive and *rational*.

TRACS*

The simulation results suggest that—by not offering an incentive to a memory update strategy—Straight TRACS is inadequate for investigating people's willingness and capacity to monitor and update changing environmental circumstances. In this section we introduce TRACS*, which provides a clear benefit for adopting an update strategy, as well as introduces additional probes of memory performance.

In designing TRACS*, we sought to create a variant of the game for which a memory update strategy clearly benefits performance. We achieved this by carefully controlling the cards dealt to the players. While cards were selected randomly, they were selected from a card space constrained by two rules. First, only pairs of face-down cards that would not have equal odds of matching the target color would be dealt. By eliminating ties, this rule eliminates the need to guess. Second, pairs were not selected if the card with the lower odds resulted in a hit, or if the card with the higher odds did not. This rule aimed to reduce the influence of luck by eliminating win-win and lose-lose situations, thus driving a wedge between the baseline and update strategies.

Figure 3 illustrates the effects of these changes. The mean score for the random agent in TRACS* remained stable, at 5.49 (out of 11) hits per game across 10,000 games. However, baseline and update scores rose to 8.22 and 10.83, respectively. Hence, our game modifications were successful in introducing a substantial benefit of the update strategy over the random and baseline strategies. Given that baseline and update strategies now yield unique performance signatures, it should be possible to determine which strategy our participants actually adopt in the game.

Our second alteration in TRACS* was procedural. In addi-

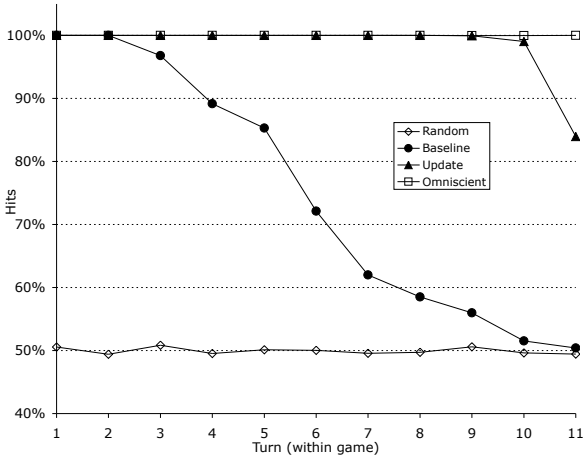

Figure 3: Simulation results for four artificial agents playing 10,000 games of modified TRACS*.

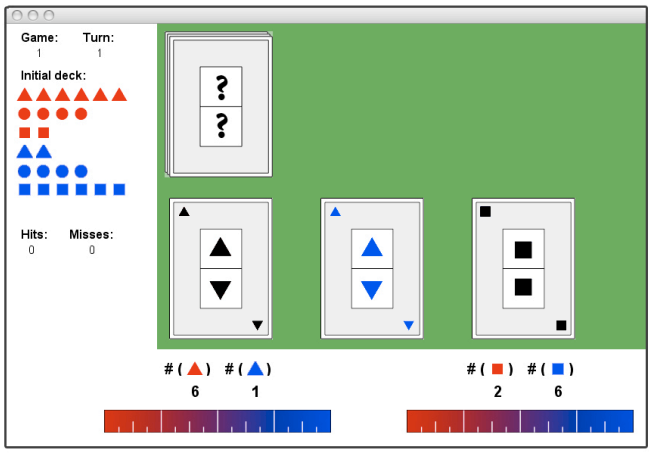

Figure 4: Screenshot of the TRACS* interface requesting odds estimates (after the completion of the recall task).

tion to using continuous color gradients to assess our participants' odds calculations, we introduced memory recall boxes to judge the accuracy of their memory. In this way we hoped to elucidate whether Burns' findings indicated an actual baseline bias, or merely just difficulty in converting accurately recalled frequencies into points along a likelihood gradient.

Experiment

Method

Twenty-five undergraduates from Rensselaer Polytechnic Institute participated in partial fulfillment of a course requirement. They ranged in age from 18 to 22 years, with an average of 19.6 years. Participants were tested individually.

The experimenter spent about ten minutes instructing each participant on the rules of original TRACS. Each participant played a total of 10 games of 11 turns each. On every turn, players had to complete the recall task, provide odds estimates, and choose a card.

On the newly added recall task participants were asked, for each face-down card, to report the number of red and blue cards of that shape which remained in the deck. Answers were typed into text boxes immediately below each face-down card. Players then estimated the odds of each face-down card being red or blue by placing a marker on a continuous color gradient. Gradients were red on the left and blue on the right, and 300 pixels wide (≈10 cm), allowing for a precision below one percent (see Figure 4 for a screenshot). Finally, participants chose a card by clicking on it. Feedback on correctness was then provided by a thumbs-up/thumbs-down image and the next turn was initiated by clicking on the feedback image.

The game was implemented in Macintosh Common Lisp 5.0 running on OS 10.2 with a 17" flat panel display set to a 1024×768 screen resolution. The initial card distribution and a hit/miss counter were shown to the left of the game window.

Results

We will assess participants' performance before turning to more detailed analyses of various error types.

Performance TRACS* allows for a straightforward correspondence between a player's awareness of the current game state and his or her outcome score. Thus, scores reliably exceeding the expected values of a simulated baseline agent would signal a memory update strategy.

On average, participants scored 9.3 hits per game with 22 out of 25 players (88%) exceeding the theoretic baseline score of 8.2 hits. This strongly suggests that memory updates contributed to task performance.

To allow for a statistical assessment of these differences, we let our simulated baseline and update agents both play the same number of games as human participants. A comparison of mean scores over the sequence of ten games per player showed that human players scored significantly more points than baseline agents [9.3>8.2, $t(26)=2.1$, $p<.001$], and significantly fewer hits than update agents [9.3<10.8, $t(25)=2.1$, $p<.001$]. Figure 5 contrasts the performance of human participants with that of simulated agents on a within-game resolution. It is obvious that human players did not perform on

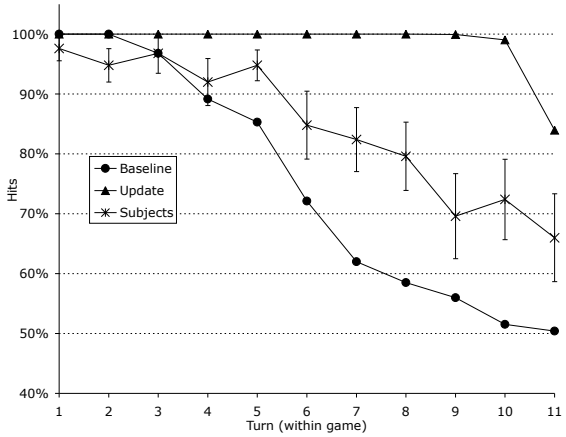

Figure 5: Participants' mean percentage of hits by turn compared to those of simulated baseline and update agents. (Error bars indicate 95% confidence intervals.)

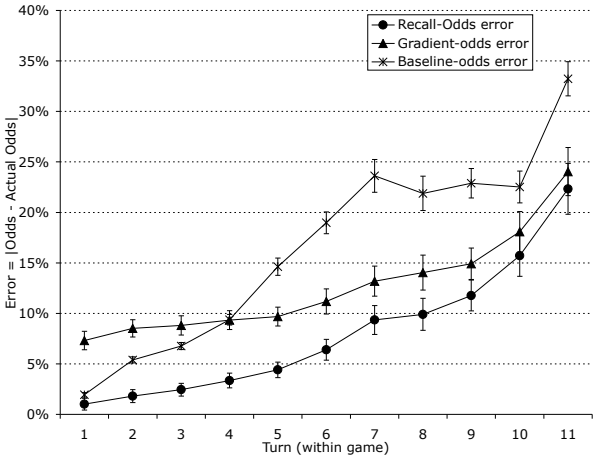

Figure 6: Average errors of odds by turn. (Error bars indicate 95% confidence intervals.)

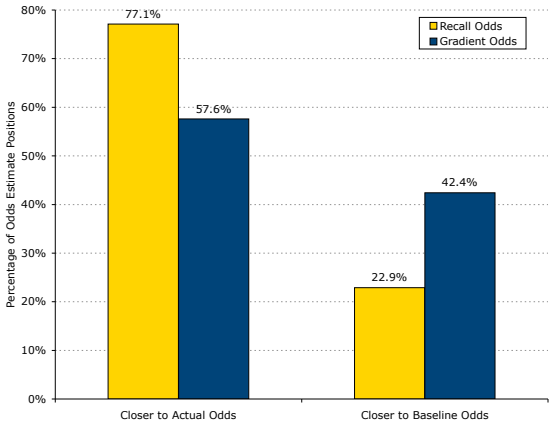

Figure 7: Percentage of odds selections closer to the baseline vs. closer to the actual value (based on n=4404 estimates).

the level of an ideal update agent, but did reliably better than a baseline agent.

To assess possible effects of learning we conducted an ANOVA with game number as a within-subjects factor. A significant main effect [$F(9,216)=3.0$, $p<.01$] indicated that players improved their scores reliably from an average of 8.8 hits in earlier to about 9.7 hits in later games. Subsequent comparisons showed that human participants outperformed a pure baseline agent in all but the initial two games.

Errors Even though human participants performed better than a baseline agent, their performance was worse than that of an ideal update agent. In this section, we examine this discrepancy by first considering erroneous frequency and likelihood estimates before assessing errors of internal consistency.

As participants estimated card frequencies as well as likelihoods we were provided with two distinct indices of memory. To allow for direct comparisons of both indices on a single scale, we converted reported frequencies into 'recall odds'. For both recall odds and likelihood estimates (as indicated on the gradient scales) we then calculated and summed up the absolute difference from the actual odds.

Figure 6 illustrates that both recall-odds and gradient-odds errors increase over the course of a game, but errors in frequency recall (with a mean of 8.0%) are significantly lower than the errors in likelihood estimates provided on gradient scales (12.6%). The third line in Figure 6 shows the mean size of the 'baseline-odds' error (16.5%) which would result if participants had adopted a baseline strategy on the given trial. Even though the mean gradient-odds error exceeded the baseline-odds error on the first three trials, the general trend indicates that participants' actual errors on both scales were lower than suggested by a baseline bias.

Taking into account the direction of deviations rather than just error magnitudes, we can also ask whether empirical recall and gradient odds are closer to the baseline or to the actual odds. Whenever the actual odds value deviates from the baseline value there are two possible attractors: Participants might specify odds closer to the baseline odds, or they might select odds closer to the actual odds. A *bias* is defined by a systematic preference. If participants—due to update failure or memory decay—were more likely to choose odds closer to the baseline than to the update value this would constitute a baseline bias. Likewise, an "update bias" could be diagnosed if participants were more likely to select odds in the vicinity of the actual value. Figure 7 shows that, in TRACS*, the evidence for an update bias clearly outweighs the evidence for a baseline bias. Participants' preference for actual values seems particularly pronounced when odds are based on recall frequencies (77.1% vs. 22.9%). In contrast, the same preference is weaker when odds estimates are measured by probability gradients (57.6% vs. 42.4%). As the baseline attractor seems to exert less gravitational pull when providing frequency estimates than when responding on a gradient scale, examining only the latter (e.g., Burns, 2002a, 2002b) might overestimate the size of a baseline bias.

All errors reported so far were deviations of empirical estimates from either true or baseline values. Our finding that participants' frequency estimates are closer to the actual values than to the initial baselines makes it implausible that participants' frequency estimates are governed by a baseline bias. At the same time, it raises questions about alternative breakdowns in performance. On the basis of our initial task analysis, the complexity of TRACS allows for a variety of non-memory related errors. In the following and final sections we consider conversion errors and errors of choice as examples of errors of internal consistency.

Due to our sequential procedure of first requiring frequency information and then asking for probability estimates, participants' responses on the likelihood gradients ought to be a direct function of recall performance. Nonetheless, people's notorious problems with probabilities can cause *conversion errors* when transforming recalled frequencies into odds on continuous scales. To assess the occurrence of such errors, we compared subjective recall odds (based on the card frequency entries of each participant and turn) with the likelihood estimates provided on the same turn. An average deviation of 6.6% indicates that this translation process was indeed non-trivial and error-prone. The magnitude of this error is striking not only as it is almost as large as the average error in fre-

quency recall (8.0%, see Figure 6), but also when considering that players reported their subjective frequencies immediately before indicating their judgment of odds and had all relevant frequencies displayed directly above the gradient scales (see Figure 4). Thus, we conclude that a large proportion of participants' error-prone responses on likelihood scales were due to errors in odds conversion.

Two curious errors of internal consistency address the relation between odds estimates and card selections. *Recall-choice errors* can be defined as instances in which the card with lower recall odds (based on the subjective card frequency estimates) is selected by the participant. Similarly, *gradient-choice errors* occur whenever the card with lower likelihood odds (based on probability estimates) is chosen.

There were 4.3% (119 out of 2750 choices) recall-choice errors, but 8.3% (229) gradient-choice errors. Given that any conflict between judgment and choice is relatively bizarre, both errors are more frequent than we would have expected. As the gradients are evaluated immediately before a choice is made, we interpret the relative size of both errors as evidence that players were more likely to base their choices on perceived frequencies than on perceived odds.

Discussion

Our first result is of a methodological nature: When creating artificial task environments to assess the scope of human rationality, the cost–benefit structure of the task must provide an incentive to display the behavior in question. Our simulation of Straight TRACS revealed that the original game provides only minimal benefits for adopting an effortful memory update strategy. This led us to re-interpret Burns' (2002a, 2002b) original finding of a 'baseline bias' as an adaptive and rational response to the properties of the task environment.

Our critique, however, does not imply that TRACS is not an interesting game and valuable research paradigm—quite to the contrary! We now believe that TRACS is both more complex and more interesting than it at first appeared. Our task analysis has suggested the need to distinguish three cognitive components: retrieving numbers of cards from memory, converting frequencies into probabilities, and mapping frequency or probability estimates to choices.

We are particularly intrigued by the errors our players made when converting natural frequency information to likelihood estimates. Players who had to provide the same information in two different formats within seconds and saw the frequencies displayed in front of them while computing probabilities still made substantial errors when coming up with simple likelihood estimates. Interestingly, our analysis of choice errors revealed that players seemed less likely to act on their inaccurate probability estimates than on their perceived frequencies even though the former just preceded their choice.

A potential caveat of our study is that by altering the cost-benefit structure of the task and assessing players' memory for card frequencies we introduced *two* changes to the original game. It is conceivable that the mere query for frequencies made the necessity to count cards more explicit, whereas it remains rather implicit in the original game. The extent to which each of our modifications contributed to the improved performance and to which a procedural task demand may have inadvertently prompted different memory strategies is an empirical question to be addressed in future studies.

Finally, the performance results of our modified version TRACS* provide a more optimistic view of the human capacity for concurrent memory updates than do previous studies. As our players were able to reliably exceed baseline performance, we conclude that the previously reported 'baseline bias' may be an artifact of the original game.

Despite our criticisms, our results agree with those of Burns (2002a, 2002b) that people are able to take baserate information into account. However, we additionally demonstrate that—when memory matters—people are also able to dynamically update their memory while being engaged in a highly demanding task.

Acknowledgments

We are grateful to Kevin Burns for allowing us to use TRACS and providing many helpful comments. In addition, we thank Christopher Myers, Bram van Heuveln and Jamie Sowder for many valuable contributions. The work reported was supported by grants from the Air Force Office of Scientific Research (AFOSR #F49620-03-1-0143), as well as the Office of Naval Research (ONR #N000140310046).

References

Anderson, J.R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Burns, K. (2001). TRACS: A tool for research on adaptive cognitive strategies: The Game of Confidence and Consequence. At www.tracsgame.com (May 2004).

Burns, K. (2002a). On Straight TRACS: A baseline bias from mental models. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 154-159. Hillsdale, NJ: Lawrence Erlbaum.

Burns, K. (2002b). Dealing with TRACS: The game of confidence and consequence. *Proceedings of the American Association for Artificial Intelligence*, Symposium on Chance Discovery.

Burns, K. (2004). Making TRACS: The diagrammatic design of a double-sided deck. *Proceedings of the 3rd International Conference on the Theory and Application of Diagrams*.

Gigerenzer, G. (2000). *Adaptive thinking. Rationality in the real world*. Oxford, UK: Oxford University Press.

Gray, W.D. (2002). Simulated task environments: The role of high-fidelity simulations, scaled worlds, synthetic environments, and microworlds in basic and applied cognitive research. *Cognitive Science Quarterly 2(2)*, 205–227.

Hess, S.M., Detweiler, M.C. and Ellis, R.D. (1999). The utility of display space in keeping track of rapidly changling information. *Human Factors 41(2)*, 257–281.

Koehler, J.J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19,* 1–53.

Newell, A., and Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Simon, H.A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*, 1–19.

Venturino, M. (1997). Interference and information organization in keeping track of continually changing information. *Human Factors, 39(4)*, 532–539.

Yntema, D.B. (1963). Keeping track of several things at once. *Human Factors 5*, 7–17.