Modeling Complex Tasks: An Individual Difference Approach

John Rehling (rehling@andrew.cmu.edu)
Marsha Lovett (lovett@cmu.edu)
Christian Lebiere (cl@cmu.edu)
Lynne Reder (reder@cmu.edu)
Baris Demiral (baris@andrew.cmu.edu)
Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

It is the usual case in cognitive modeling that a model's output is compared to the average of a number of subjects, in which case the enterprise of modeling is apparently to capture the behavior of the typical individual. Our approach is to administer two simple tasks to each subject, using performance on those tasks as measures of individual ability. Those measures are then used as the values for parameters in an ACT-R model of a more complex task, so that the model can predict individual performance on that task.

Introduction

Work in cognitive modeling, when it seeks validation in the performance of human subjects, is almost unanimously concerned with the average performance of many subjects. For many purposes, however, it is desirable to be able to model or predict *individual* performance. We present here the first work to use a fine-grained cognitive model to predict individual performance in a complex task.

The ACT-R architecture, the basis of a great deal of work in cognitive modeling, has a detailed, well-developed theory of cognition – perception, learning, performance, and so on (Anderson and Lebiere, 1998). The architecture by necessity contains a number of parameters that can be used to fix levels of performance in, e.g., memory, to realistic levels. The ACT-R community has by custom sought universal values for these parameters wherever possible, finding values work across tasks, optimizing how well the model fits the data of the average subject. These parameters are each *meaningful*, each parameter determining the model's behavior in one specific way. For example, there is a parameter called W that determines the sum of the activations of all the pieces of information that may be retrieved at any point in time. It therefore controls the model's working memory capacity. Extensive empirical work in ACT-R modeling (again, of the kind where the model was meant to predict the average subject) found that a W value of 1.0 produces very good fits with subject data.

It was later postulated, however, that the W parameter could be meaningfully varied in order to model individual differences in working memory capacity (Lovett, Reder, & Lebiere, 1999). This was later demonstrated empirically by using individual performance in one simple memory task to measure the W value that best fit the individual's performance. The diagnostic memory task is called MODS, or the modified digit span task. In each MODS trial, subjects are presented strings of digits to be read aloud in synchrony with a metronome beat and are required to remember the final digit from each string for later recall. After a certain number of digit strings are thus presented, a recall prompt cues the subject to report the memory digits in the order they were presented. Each subject's MODS score was used to estimate their individual W value, which was then plugged into an ACT-R model of a separate working memory task, and the model output was used to predict *individual* data on that second task (Daily, Lovett, & Reder, 2001).

Previous and concurrent work by other groups suggested a number of positive characteristics that might be combined into a single, more powerful methodology. ACT-R parameters had been manipulated (Taatgen, 2001) in a model of individual differences in learning, although the "individuals" in that work were simulated, and corresponded to *types* of individuals, not to actual subjects. In that work, performance in a complex task was related to individual difference parameters across the simulated individuals. Earlier work in modeling also accounted for relationships between ability in one task and ability in another, but making assessments on the group, not individual, level (Just, Carpenter, and Shell, 1990). A complementary approach measures individual performance on complex tasks, and utilizes statistical methods such as intercorrelation matrices, allowing predictions of individual performance on one task based upon measurements of performance on other tasks in the matrices, making no use of any particular theory of cognition (for example, Ackerman and Kanfer, 1993).

All of this previous work, we felt, pointed towards a methodology that combined a number of positive features from these complementary approaches into a

single, more comprehensive modeling paradigm In the methodology we envisioned, one or more simple tasks could be administered to an individual, allowing us to estimate their individual parameters; then, by plugging the individual's parameter values into the ACT-R architecture, we could predict the individual's performance on any task for which an ACT-R model exists Because ACT-R models produce predictions with a grain size of tens to hundreds of milliseconds, this provides us with a *detailed* model of individual performance, offering the potential for predictions on the trial level, or predictions of novel measures of performance that emerge from lower-level detail – potentially allowing predictions of almost any measure that can be made of subjects. Because our approach builds atop the platform of the rich ACT-R theory, it is realistic to expect that these individualized model runs will be somewhat meaningful in their details, not just a way to arrive at a final, aggregate performance metric of some kind

In order to take the next step beyond the Daily, Lovett, & Reder study that involved only two simple memory tasks, we decided to pick a more complex, interactive task In order to capture a broader spectrum of individual differences, we chose to measure two parameters per subject: the W parameter as well as a measure of perceptual and motor ability, henceforth referred to as P/M This is not a part of the standard ACT-R architecture, but seemed to be an important kind of individual variation Thus far, we have used only one parameter, which represents as though they were one individual perception and motor speed. We allow that those may covary freely among individuals, but we have so far had success using the one parameter alone for this

The AMBR Task

Given the preceding considerations, we chose as our more-complex task the AMBR simulation, an air traffic control task that already had a foundation as a test bed for cognitive models in a project organized by the Air Force Research Laboratory (Gluck and Pew, 2001) This task already had an ACT-R model implemented (Lebiere, Anderson, and Bothell, 2001), which not only facilitated our project, but also provided a gauge of the modularity of our approach; ideally, we would be able to plug parameters into this off-the-shelf model and obtain good results without modifying it in any other way

The task places the subject in the role of an air traffic controller whose job is to process aircraft (AC) as they enter and leave the airspace zone, central in the simulated radar display, for which he or she is responsible This primarily consists of issuing, via a graphical interface, two commands to an AC as it enters one's zone from a neighboring zone of airspace, and issuing two commands to an AC as it departs for another zone The same AC must thereby be issued a total of four commands if it passes into and subsequently out of the central zone during a scenario In some cases, the AC will only enter the central zone, or only depart the central zone, during the duration of a scenario, in which case that AC will require only a total of two commands In addition, a fifth type of command is required if an AC requests a speed change, which requires the subject to make a trivial judgment as to whether or not the AC is on course to catch, from behind, any other AC; if so, the speed change request should be denied, and otherwise, it should be accepted AC arrivals can be detected both from the radar display and from text messages appearing in windows to the side of the display. Speed change requests can be cued only via text messages. The departure of an AC from the central zone can be detected only via the radar display. Under the assumption that AC are at different altitudes, however, collisions cannot take place in this simulation, nor do AC land or take off in the simulation. The subject is scored based on issuing all commands in a timely fashion that permits AC to move freely without ever reaching the border of the central zone while still awaiting one of the required commands. If an AC does reach the zone border without having received all necessary commands, it will go into a hold, thereby turning the AC red in the display, halting the AC's motion, and penalizing the subject 1 point. The score at the end of the run is the sum of the errors the subject makes, lower score thereby signifying better performance. Subjects were also penalized for making interface errors of the sort that the model never made. Subject and model performance levels can thereby be compared on the basis of hold errors. A static image of the display is visible in Figure 1.

We were required to modify one aspect of the AMBR task in order to eliminate uncontrolled strategic variation among the subjects. AMBR's original implementation has a more baroque scoring system where some errors lead to penalties of 1 point and other errors up to 50 points. In response to that scoring system, some subjects tried to avoid all errors while other subjects opportunistically allowed low-penalty errors when that helped them avoid any occurrences of high-penalty errors. That strategic variation was noticed only when some data had been collected; this is an indication of the subtle difficulties that can arise when modeling tasks at the level of complexity of AMBR. The difficulty of producing a suitably correct Cognitive Task Analysis was roundly reported by the four cognitive modeling groups involved in the AFRL's AMBR modeling project. Unconstrained by the need to need to coordinate with other groups, we changed the task.

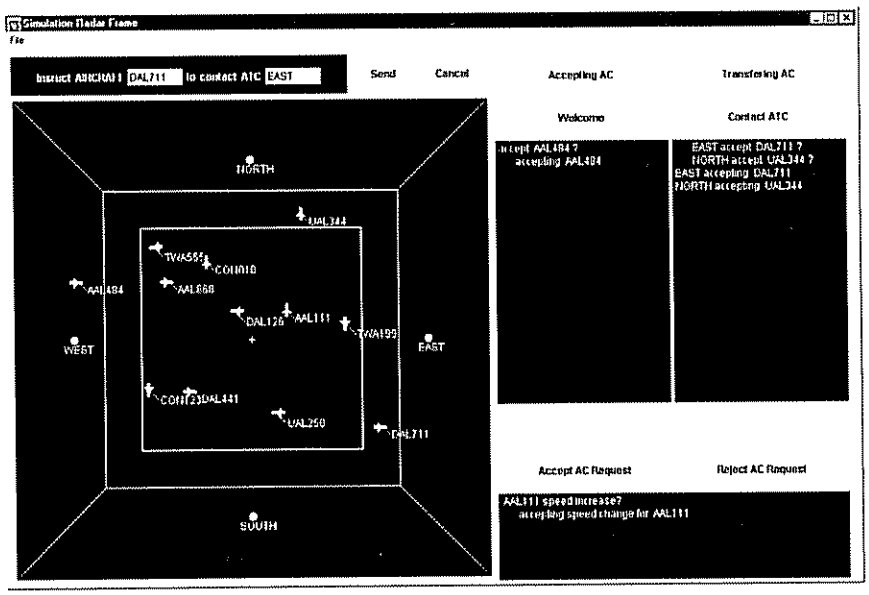

Figure 1: The AMBR Display

Sources of individual difference

Having introduced AMBR as our complex task, we acknowledge that the term "complex" is a relative one, and in seeking a complex task for our work, we were actually seeking an appropriate kind of complexity We distinguish between distinct kinds of individual difference factors, postulating that **architectural differences** are those differences that pertain to relatively permanent characteristics of the individual, not shaped by particular episodes in the individual's experience (We make no claims about how development shapes architectural differences throughout an individual's life.) **Knowledge-based differences**, on the other hand, can arise through specific instances of learning declarative information; the state of an individual's knowledge can only be described (or tested) in a very expansive manner, and this is not our enterprise A third type of individual difference, **strategic differences**, could be broken down into either of the two previously mentioned types It is not our goal to measure the encyclopedic total of an individual's knowledge, but we do anticipate that certain differences in how an individual chooses a strategy for a given task will depend upon and emerge from architectural differences Cases where we can predict strategic differences based upon architectural differences will serve to validate our approach We recognized that we would have trouble, however, with any task that invited strategic variation between subjects that could not be predicted from architectural differences In such a case, our individual-difference approach would risk the same pitfalls that a non-individual-approach can lead to when subject strategies vary (Newell, 1973; Siegler, 1987)

Initial results

In two distinct experiments, with two sets of subjects, we applied the methodology of administering initial tests to measure the W and P/M parameters The P/M parameter was actually calculated based upon the speed of mouse clicks in the AMBR training Our procedures for calculating the parameters produce values of W and P/M which both have population means of about 1 0 and standard deviations of about 0 2 (for Carnegie Mellon undergraduates) High W means better working memory capacity, while high P/M means slower perceptual and motor responses – it is a multiplier, so that P/M = 1 2 means responses 20% slower than average) Therefore, where we find significant effects, W correlates negatively with error counts and P/M positively

Subjects were trained on the AMBR task until they understood it quite thoroughly, and then participated in a number of AMBR scenarios, the data from which we compared to individualized model runs for each subject Experiment 1 featured 10 AMBR scenarios, each 9 minutes long, and alternating between very easy and very difficult Experiment 2 had 9 scenarios, each 4 5 minutes long, varying evenly along a continuum in terms of difficulty from easy to difficult As an informal measure of difficulty, we have taken the number of AC per scenario times the average speed of those AC, divided by the scenario length. Using the idiosyncratic units of our simulations, Experiment 1 scenarios had difficulty ratings of 26 (easy) and about 180 (hard). Experiment 2 scenarios ranged in difficulty from 40 to 200

It is instructive to note the analysis that would be performed if this were not an individual difference study. Aggregate group performance, measured in hold errors, as a function of scenarios was predicted well by the aggregate model runs (Experiment 1: r = 0.975; Experiment 2: 0.929). This was almost the same analysis, from the very same ACT-R model, presented in Lebiere, Anderson, and Bothell (2001), and used to argue for a good fit between subjects and model.

The model correctly predicted that the AMBR task, as originally conceived, is more sensitive to variation in P/M than in W. This is seen clearly in the correlations of subject holds with P/M (r = 0.658) and W (r = -0.266). Not only can the model be used to generate predictions for specific subjects, but it can also be used to probe the effects of one parameter by varying that parameter while holding the other one neutral (at the population mean of 1.0). (Note that holding one parameter fixed while varying another among the subjects is a very difficult practical matter.) This use of the model shows a strikingly greater effect upon holds from P/M than W. This is in agreement with data on the actual air traffic controller task (which, it should be noted, has several distinct differences from the AMBR task, not the least of which being that it involves voice communication, not a graphical user interface alone), which documents that only a small number of errors are due to memory failures (Billings and Cheaney, 1981).

Studies of AMBR traces reveal that the reason for this is that hold errors are primarily an outgrowth of time pressure when the time demands on a subject exceed the time that is available. For 3 of the 5 types of command in AMBR, the subject is shown the name of an AC in the text cue, and must click on the AC as part of the subsequent action sequence. Memory becomes a factor in AMBR performance primarily in that if a subject cannot remember the location of an AC based upon its name, then the display must be searched for the AC. This turns out to be a small factor because visual search is fast – slower than memory, perhaps, but the difference is on the order of a fraction of a second, while clicking in a command sequence takes several seconds whether the AC location is remembered or not. W, then, is logically a small factor in the original AMBR task, and for a small portion of the variance.

Designing for science

In order to improve upon the studies described above, we designed a follow-up study that modified the scenario difficulty, the measures of performance that we used, and even the task itself. It was obviously necessary to decrease scenario difficulty into the range for which the model produced a good fit to subject data. This also allowed us to use one performance measure that is more sensitive than hold errors – the reaction time between an action's cue and the subject's response to that cue (we used the time for an action sequence to end, meaning the third or fourth click in all). In order to emphasize the W effect relative to the P/M effect (since cognitive modeling, and not motor/kinesthetic modeling is our chief interest), we modified the task so as to create a greater penalty for failures in recall. We did this by removing AC names from the display by default, and showed the name of an AC only when the subject clicked on the AC. Moreover, only one AC name could be seen at a time, and this would appear after a delay. This change meant that the speedy visual searches of the earlier experiments would be impossible, and any failure to recall an AC's location would entail an excruciatingly slow manual search. This task modification also had the merit of giving us data on searches, and let us emphasize a performance measure that calculated what on what proportion of commands a subject found the correct AC on the first click. In other ways, Experiment 3 was similar to Experiments 1 and 2. Each subject was to participate in 5 AMBR scenarios that were easy – hold errors confound reaction time, so we needed them to be fairly rare in order to use RT as a performance measure. In the units of scenario difficulty mentioned earlier, all Experiment 3 scenarios gauged 17 or lower.

Before the study began, we ran the model, which was revised to allow for the task modification involving name-hiding, on the Experiment 3 scenarios, and it seemed not to work correctly. Instead of performing manual searches for AC names, it would *guess* which AC it was looking for and click through the entire action sequence without bothering to verify that it had clicked the right AC. While work on the model, to fix this "problem" was underway, the first subjects ran in the experiment. They behaved the same way. We had set the delay that one must wait, after clicking on an AC, for its name to appear, too long, and subjects preferred to hope that they had guessed right correctly rather than perform the laborious verification process. ACT-R came to the same conclusion based upon the undesirably large cost associated with clicks that required several seconds before the desired consequence took place. We modified the task again, shortening the delay before the name appeared, and both the model and the subjects performed manual search in the way we had hoped. This demonstrates one possible application of our approach – tasks (experimental or otherwise) can be *designed* with the model's predictions taken as a serious indicator of subject performance, individual or otherwise.

Experiment 3 produced the subject characteristics we had sought. Our three measures of individual

performance correlated significantly with W (Holds: r = -0.444; RT: r = -0.314; First-clicks: r = 0.314) P/M had about as large an impact on performance (Holds: r = 0.508; RT: r = 0.485; First-clicks: r = -0.172), but, as we desired, it did not dominate as in the first two experiments.

The result most central to our intent was the prediction of individual performance with model output (Holds: r = 0.461; RT: r = 0.436; First-clicks: r = 0.406) These correlations are distinctly less than what is often possible when averaging multiple model runs against the average of many subjects, but are very much in line with the kinds of correlations found in task intercorrelation matrix approaches (Ackerman and Kanfer, 1993; Joslyn and Hunt, 1998)

To demonstrate the possibility of precise, instance-level predictions, we looked at model predictions across all three experiments, as to whether or not, for each scenario, an individual subject would commit at least one hold error The model predicted correctly 91.7% of the time, as detailed below in Table 1

	Subject scenarios with errors	**Subject scenarios with no errors**
Model scenarios with errors	205	4
Model scenarios without errors	21	70

Table 1: Prediction of Error Situations

Future directions

For a variety of goals, both applied and scientific, it is and will be desirable to be able to predict individual performance on a fine-grained level It seems certain that the methodology we are exploring will be expanded upon and utilized for such applications in the future. At present, it is possible to point to the range and extent of our successes and note the particular difficulties that individual difference modeling entails

One avenue to explore is to involve a larger number of individual difference factors ACT-R has many parameters built into it, and future work may be able to predict individual performance more accurately by making use of pre-tests besides the two we now use

Because our model is fine-grained, it permits many measures of performance, on the subject, scenario, command, or click level Ways in which the model fits, or alternately does not fit, subject data highlights many areas where future work is required For example, we have observed in the subject data from Experiment 3 some phenomena of interest that the model does not predict These include a correlation between higher W and the frequency with which a subject completes a sequence of command clicks *without* waiting for the AC name to appear We believe that we can capture this with additional refinements to the model, taking advantage of ACT-R's utility-learning mechanisms A second discrepancy between the subject data and the model predictions are that the model does not recall AC locations as well as the subjects do, and we believe that this stipulates that the model should include rehearsals of AC location between the time that information is learned and when it is needed. A third difference is that subjects often respond to Welcome commands, which are always the second of a pair of commands regarding a given AC, much faster than the model does. In fact, some subjects respond much faster than other subjects in this regard, and it is clear that strategic variation has intruded into our study – something that is difficult to prevent absolutely with a task of AMBR's complexity In upcoming experiments, we will try to instruct all subjects to anticipate Welcome commands when they can, and will change the model so that it does so as well

Subject phenomena that are not captured by the model, we believe, stem from the problem of deriving a valid Cognitive Task Analysis, which is known to be difficult for a novel, complex task It is striking how much simpler AMBR is than many tasks (for example, *real* air traffic control), and yet how challenging it is to model it precisely It has not only been a challenging task to which to extend the individual difference methodology from memory to more complex tasks; it is also at the right level of complexity for the next stages of work as we try to model it still more accurately and over a variety of task modifications.

Acknowledgments

This research was supported in large part by ONR Grant N00014-02-10020 Thanks to Susan Chipman for her support

References

Ackerman, P L & Kanfer, R. (1993) Integrating laboratory and field study for improving selection: Development of a battery for predicting air traffic controller success *Journal of Applied Psychology, 78(3)*, 413-432

Anderson, J R. & Lebiere, C. (1998) *The atomic components of thought* Mahwah, NJ: Erlbaum

Billings, C. and Cheaney, E. (1981) The information transfer problem: summary and comments *Information Transfer Problems in the Aviation System* NASA Technical Paper 1875 NASA, California

Carpenter, P.A., Just, M A , & Shell, P (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test *Psychological Review*, 97 (3), 404-431

Daily, L. Z , Lovett, M C & Reder, L. M (2001). Modeling individual differences in working memory performance: A source activation account *Cognitive Science*, 25, 315-353

Gluck K A , & Pew, R W (2001) Overview of the Agent-based Modeling and Behavior Representation (AMBR) model comparison project *Proceedings of the 10th Computer Generated Forces and Behavior Representation Conference*, Orlando, FL.

Joslyn, S & Hunt, E (1998). Evaluating individual differences in response to time-pressure situations. *Journal of Experimental Psychology Applied, 4(1)*, 16-43

Lebiere, C., Anderson, J. R., & Bothell, D (2001). Multi-tasking and cognitive workload in an ACT-R model of a simplified air traffic control task *Proceedings of the 10th Computer Generated Forces and Behavior Representation Conference*, Norfolk, VA.

Lovett, M C , Reder, L. M , & Lebiere, C (1999). Modeling individual differences in a digit working memory task. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 460-465). Mahwah, NJ: Erlbaum.

Newell, A (1973). *You can't play 20 Questions with nature and win: Projective comments on the papers of this symposium* In W Chase (Ed), Visual information processing New York: Academic Press.

Siegler, R S (1987) The perils of averaging data over strategies: An example of children's addition. *Journal of Experimental Psychology General*, 106, 250-264.

Taatgen, N A (1999) A model of learning task-specific knowledge for a new task In *Proceedings of the twenty-first annual conference of the Cognitive Science Society* (pp 730-735) Mahwah, NJ: Erlbaum