In Proceedings of ICCM - 2007- Eighth International Conference on Cognitive Modeling. 299 - 304. Oxford, UK: Taylor & Francis/Psychology Press.

Modeling the Range of Performance on the Serial Subtraction Task

Frank E. Ritter[1] (frank.ritter@psu.edu), Michael Schoelles[3], Laura Cousino Klein[2], and Sue E. Kase[1]

[1]College of Information Sciences and Technology, and [2]Biobehavioral Health Department
The Pennsylvania State University, University Park, PA 16802 USA
[3]Cognitive Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180 USA

Abstract

We present a model of serial subtraction, a task where subjects repeatedly subtract a 1- or 2-digit number from a 4-digit number. The model performs 4 min. blocks of these subtractions like subjects do. The current model replicates part of the pace and % correct for group data. Because performance on this task varies widely between subjects, we explore what it means to match the data distribution. We find that our model represents individual subjects better than group means. We can start to model a distribution of performance and illustrate some of what this approach will entail.

Introduction

Serial subtraction, repeatedly subtracting a 1- or 2-digit number from a 4 digit number is part of the Trier Social Stressor Task (TSST, Kirschbaum, Pirke, & Hellhammer, 1993). This is an interesting task for two reasons. One reason is that it has been used over 100 times in published articles to study the effects of stress on physiology (e.g., Kudlielka, Buske-Kirschbaum, Hellhammer, & Kirschbaum, 2004; Nater et al., 2006; Taylor et al., 2006; Tomaka, Blascovich, Kelsey, & Leitten, 1993). It is a cognitive task used to cause stress, but we don't know how it's performed—there is only one report on how well it is performed (Tomaka, Blascovich, Kelsey, & Leitten, 1993), and this report only provides data on one 4-min. block.

The second reason it is interesting is that subtraction is an interesting task in its own right and as a component task to many other tasks. It involves many cognitive mechanisms making it a good task to study cognition, not just the biobehavioral effects of laboratory stress. Real world tasks that use subtraction include air traffic control, navigation, and piloting the wide range of vehicles that use angular directions.

It would be useful to have a cognitively plausible model of performance of subtraction. This model would serve as an explanation and summary of task performance, helping to summarize regularities, and a model would also be the starting point of a theory of how cognition changes with stress. Because the task requires not only executive control and memory but interaction with the verbal system as well, a model will be able to quantify the constraints that these subsystems of cognition impose on the task. These requirements suggest that the model be constructed on an embodied cognitive architecture (Anderson, in press).

Previous work with an earlier model has shown that the general pattern of high level results (i.e., number of attempts per 4-min. block and percent correct) with serial subtraction can be predicted (Ritter, Reifers, Klein, Quigley, & Schoelles, 2004), and we have used this approach to describe how popular theories of stress could influence performance on this task (Ritter, Reifers, Schoelles, & Klein, 2007). The next steps, presented here, are to create more detailed predictions of performance and compare these predictions to more detailed subtraction performance data than has been previously presented.

The remainder of this paper presents a serial subtraction experiment, the architecture and model, subtraction data, and a comparison of the model with human data. The model's predictions match the individual data fairly well, and provide lessons for understanding how serial subtraction is performed. The model-data comparison also makes suggestions for the further development of cognitive architectures.

The Serial Subtraction Experimental Data

As part of a larger project on the biobehavioral effects of stress in men and women, serial subtraction was administered as part of the TSST. Several aspects of serial subtraction performance were recorded. We present several of them here as an initial summary of performance on serial subtraction. They are taken from a more complete report (Ritter, Bennett, & Klein, 2006).

Subjects

Thirty-six healthy women and 20 men, 18-30 years of age (μ=21.1) were recruited to participate in a study examining hormonal responses to stress.

Method

All subjects participated in the same protocol, which consisted of a baseline rest period, the TSST challenge period (approximately 30 min.), and a recovery period.

Following informed consent and a baseline rest period, participants were asked to complete the TSST which consisted of: (a) preparing a 3.5-min. speech on a personal failure, which they were told would be recorded for later observation, and then (b) completing two blocks of serial subtraction across a 15-min. period. The first subtraction set required counting backwards from a 4-digit number by 7's; the second set required counting backwards by 13's.

Subjects' serial subtraction answers were corrected against a list of answers from the starting 4-digit number. When an incorrect answer was given, the subject was told to "Start over at <the last correct number>". At 2 min. into each 4-min. session, subjects were told that "2-minutes remain, you need to hurry up".

Performance on the first block of 7's and first block of 13's were recorded on the experimenter's scoring sheets. Part way through the study a mark to indicate where the 2-min. warning occurred was added to measure pace of the subtractions. Subjects were paid $30 for their time at the end of the study.

Results

All 56 subjects completed the task. Table 1 shows the subtraction rates. Overall performance was generally accurate. The proportion correct was not different across problem types (t(56)=1.7, ns).

Table 1. Serial subtraction performance on 4-min. blocks of 7's and 13's, means, (SD), and [ranges].

(N=56)	7's	13's
Attempts	47.0 (17.1) [8-106]	36.3 (15.1) [9-78]
%Correct	82% (14) [43-100]	78% (17) [31-100]

These results are fairly comparable to Tomaka et al.'s (1993) data of 61 attempts per block of 7's for their subjects that saw the task as challenging and 46 for their subjects who saw the task as threatening. While we do not know the variance in Tomaka's data, we can compare it to this data assuming that the variance in each case is equivalent. If we do so, for number of attempts and number correct there is not a reliable difference between this data and Tomaka et al.'s (1994) threatened condition $t(36)<1$, however, there is a reliable difference between this data and his challenged condition $t(36) > 4$, $p<0.05$. There is also a reliable difference for proportion correct, with Tomaka's subjects being correct more often (91% and 92% correct, respectively).

A wide range of performance is found. Figure 1 shows that for the first block of 7's the number of attempts ranged from 8 to 106 attempts, and the number correct and error rates had similar variance. The second block, the 13's, had similar variance. The range of these scores suggests more individual variability than implied by Tomaka et al.'s values or the means in Table 1.

Error rates by sub-blocks were computed for subjects where the scoring sheet was marked with the location of the 2-min. warning. These scores are shown in Table 2. Line 3 in the table shows that subjects made many more errors in the second half of the experiment than in the first half (e.g., 6% of the 7's errors were in the first half, 94% in the second). This trend appeared to be consistent across problem types: On the 7's problems, 33 of the 34 subjects increased their errors in the second sub-block; on the 13's, 30 subjects increased their errors.

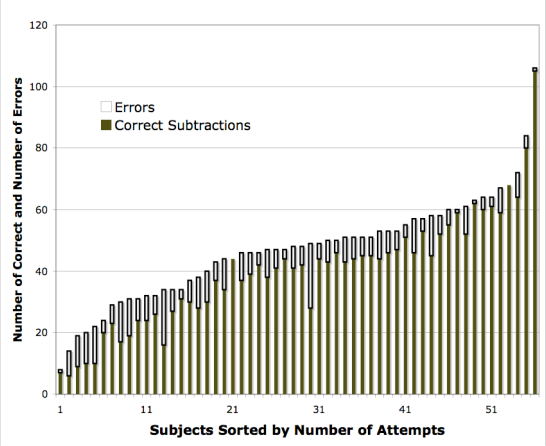

Figure 1. Histogram of attempts and errors for the 7's block.

Table 2. Serial subtraction performance before and after the 2-min. warning.

	7's		13's	
(N=34)	**Pre-2-min.**	**Post-2-min.**	**Pre-2-min.**	**Post-2-min.**
Errors	0.94 (1.3)	6.65 (3.7)	1.15 (1.9)	6.58 (4.1)
Min/max	0/5	1/20	0/8	1/26
Error %	6 (12)	94	13 (21)	87

These errors could have occurred either because of fatigue, the cumulative effects of stress, memory effects such as proactive interference, or perhaps due to the interruption. Or, it could be due to a combination of these effects. This effect is not surprising, in that many theories of stress predict that one starts out good and gets worse as time progresses (Ritter et al., 2007). More fine-grained human data, which we are preparing from another study, will be required to see where and how the increase in errors occurs.

Summary of Study

The data from this study extend the details of how serial subtraction is performed. We provide further details on this task, including means, SDs, and ranges on subtraction attempts, correct subtractions, and errors by sub-block. This study also provided data on another problem size (13's). The 13's problems appear to be slightly more difficult than the 7's, which might be expected (13's problems have about 38% more simple subtractions, and this ratio here is 30%).

The results confirm the rate of subtractions by 7's for 4-min. blocks previously found, but the rates found here are slightly slower than Tomaka et al. (1993, exp. 2) found. There may be several reasons for the lower number of subtractions per 4-min. block here than in the Tomaka et al. study. The subjects in this study may have been in a more threatening condition. While Tomaka et al.'s subjects were connected to an EKG, subjects in this study were connected to a blood pressure machine, had an indwelling catheter in their arm, and their subtraction attempts were preceded by a talk to a video camera on "an embarrassing incident." Tomaka's subjects had a longer and more relaxing break between sessions than did these subjects (5 min. rest vs. a word problem set that took about 4 min.).

The results show a trend to increasing errors with time. Nearly all subjects made most of their errors in the second half of the tasks. While we cannot see exactly where the errors occurred, it does appear that either the warning or the time on task eventually leads to errors.

The Serial Subtraction Model

ACT-R 6.0 (Anderson et al., 2004) is a useful architecture to model this task for three reasons: (a) It provides a subsymbolic level to implement changes in processing; (b) it permits the parallel execution of the verbal system with the control and memory systems, which appears to be important for this task; and (c) ACT-R has been used for other models of addition and subtraction developed by other researchers. Therefore, the representation of integers and mathematical rules can be transferred from these to other math models.

Overview of ACT-R

ACT-R is a two layer modular architecture based on the production system framework. One layer contains symbolic representations and has a serial flow in that only one production can fire at a time. The second layer is a sub-symbolic layer with numeric quantities as representations that are the result of computations performed as if they were executed in parallel.

Figure 2 shows ACT-R's modular architecture. The ACT-R modules communicate through buffers, which can hold a single copy of a declarative memory chunk. The default set of modules can be partitioned into Perceptual, Motor, Control, Memory and Representation Modules. The model presented in this paper exercises the Declarative, Procedural, Goal, Imaginal, and Speech modules. This section describes the details of these modules at the level necessary for understanding our model.

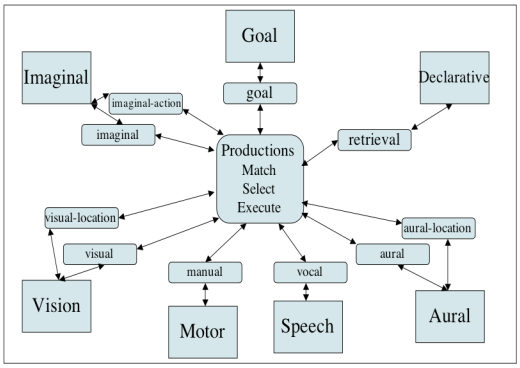

Figure 2. The ACT-R 6 Architecture.

The Declarative Module and the Retrieval buffer make up the declarative memory process. Declarative Memory contains chunks that are typed slot-value objects representing facts. At the sub-symbolic level chunks have a numerical activation value, which quantifies memory operations. Activation in this model is determined by the recency and frequency of use of the chunk plus a component that reflects retrieval system noise. Productions request the retrieval of the chunk from Declarative Memory that has the highest activation among all chunks that match a specified retrieval pattern above a retrieval threshold. Activation represents the degree to which the chunks have been learned and decays over time. Chunks are either created

initially or are created during processing. For initially created chunks, the activation can be set as if the chunk had been created at some date in the past and had been previously used. Chunks created during processing are created with a specified base level of activation.

The Procedural Module contains Procedural Memory that consists of condition-action rules (productions). The productions represent procedural knowledge. At each cycle, the conditional constraints specified in the productions are matched against the contents of the buffers. All matching productions are entered into the Conflict Set. The production to execute is determined at the sub-symbolic level by calculating a utility value for each matched production. The production with the highest utility is executed, which consists of performing the operations specified in its actions.

The Imaginal Module buffer implements a problem representation capability. In the Serial Subtraction Model the Imaginal Buffer holds the current 4-digit number being operated on (i.e., the minuend) and the subtrahend. The Goal module and goal buffer implement control of task execution by manipulation of a state slot.

The Speech Module and Buffer speak the result of each subtraction. The rate of speech is a parameter that specifies the rate in seconds per syllable.

The model

Our model of serial subtraction described starts with a main goal to perform a subtraction and a borrow goal to perform the borrow operation when needed. Both types of goal chunks contain a state slot, the current column indicator, and the current subtrahend (i.e., the number being subtracted). The current problem is maintained in the Imaginal Buffer. This buffer is updated as the subtraction is being performed.

The model starts out with an integer minuend (i.e., the number being subtracted from) of 4-digits. All numbers in the model are chunks of type integer with a slot that holds the number. The model also contains subtraction and addition fact chunks whose slots are the integer chunks described above. This representation of the integers and arithmetic facts has been used in many ACT-R arithmetic models and therefore is a good example of reuse.

The model outputs the answer by speaking the 4-digit result. It has two strategies for answering. The *calc-and-speak* strategy speaks the result in parallel with the calculation of the answer. That is, if the current problem is subtract 7 from 8195 the model would have the speech module speak "eighty one" while the operation of subtracting chunk seven from chunk five was being performed. The other strategy is a basic strategy where the answer is spoken only after the entire subtraction has been performed. All results here are obtained using the calc-and-speak strategy.

The model determines if a borrow operation is required by trying to retrieve a comparison fact that has two slots, a greater slot containing the minuend and a lesser slot containing the subtrahend. If the fact is successfully retrieved then no borrow is necessary, otherwise a borrow subgoal is created and executed.

Borrowing is performed by retrieving the addition fact that represents adding ten to the minuend. The subtraction fact with the larger minuend is retrieved. The model then moves right one column by retrieving a next-column fact using the current column value as the cue. If this retrieval fails then there are no more columns so the borrow subgoal returns back to the main task goal. If there is a next column and its value is not zero then one is subtracted from it by retrieval of a subtraction fact. If the value was 0 then the problem is rewritten in the Imaginal Buffer with a 9 and the model moves to the next column and repeats the steps discussed above, returning to the main task when there are no more columns. If the answer is incorrect, the problem is reset to the last correct answer.

In the main task when the subtraction is complete, the problem is rewritten in the Imaginal buffer and the model speaks the answer using one of the speaking strategies.

There appear to be three important parameters for this model. The rate that the model speaks is controlled by the syllables-per-second parameter (SYL). The retrieval time is controlled by the base level constant (BLC) and decay parameters. The error rate for retrievals in this model is due to the activation noise parameter (ANS). In collecting the model data these parameters (except the decay parameter) were varied to produce outcomes discussed in the results section.

The Model and Data: Matching the Range of Human Performance

The model's average performance with values of SYL=0.15 s/syllable (ACT-R default), ANS=0.1, and BLC=1 (ACT-R default) was 77 attempts with 83% correct, with no values below 68 attempts. This does not match the distribution of human data. Thus, we started to search for

parameter values and parameter value sets to match our subjects' performance.

Figure 3 shows a summary of a parameter sweep on these parameters (ANS: 0.01 to 0.71 by 0.035, SYL: 0.01 to 0.68 by 0.035, and BLC: 1 to 1.95 by 0.05, 1 run/value, 8,000 total runs). The plot shows, for ranges of parameter values, how many runs (across the sets of other parameter values) were within the range of subject performance for number of attempts and % correct. The lines on the left are for SYL and ANS. These lines show that very fast speaking rates are too fast, but otherwise there appear to be a relatively wide range of acceptable values. The other line shows that as ANS increases, the percentage of runs that are within the range of human performance increases as well, and then drops off. The line to the far right is for BLC. It shows that BLC=1.6 (and 1.85) led to a local maximum number of runs that were within our subject range. (The percentages of useful model runs in Figure 3 appear to be somewhat low because, for example, the point at 1.5=BLC contains all the values for SYL and ANS, including quite poor combinations.)

Table 3 and Figure 4 thus show the distribution of performance with the peaks of the parameters tested in Figure 3 (SYL=0.15, ANS=0.38, BLC=1.6 and 1.85). These two distributions have more runs that are within the range of performance by the subjects (which is shown in Figure 3), but the resulting distributions of performance shown in Figure 4 are less like the subjects' performance than the default parameters.

Figure 4 suggests that a distribution of parameters is likely to be more representative of the range of subject performance. The settings of the model shown in Figure 4 appear to match individual subjects (or small sets of subjects) much better than they match the whole distribution. We believe this is because the subjects have different speaking rates, different resources (e.g., working memory and knowledge), different appraisals of the task (and thus different noise and anxiety settings), or other differences we have not yet explored.

Table 3. Performance by the model on 4-min. blocks of 7's and 13's, with SD and ranges for SYL=0.15, ANS=0.38, and BLC=1.85).

(N=100)	7's	13's
Attempts	58.3 (2.2) [56-68]	44.3 (1.95) [39-50]
%Correct	65.7 (21.5) [2-84]	85.3 (13.1) [27-98]

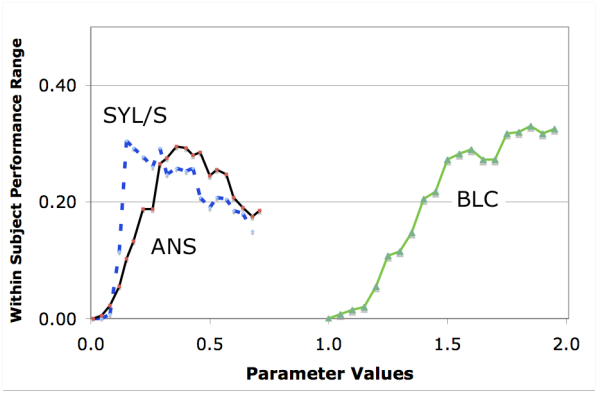

Figure 3. Summary of performance within the range of subject performance (for attempts and % correct) for 7's problems. (Each point is 400 runs.)

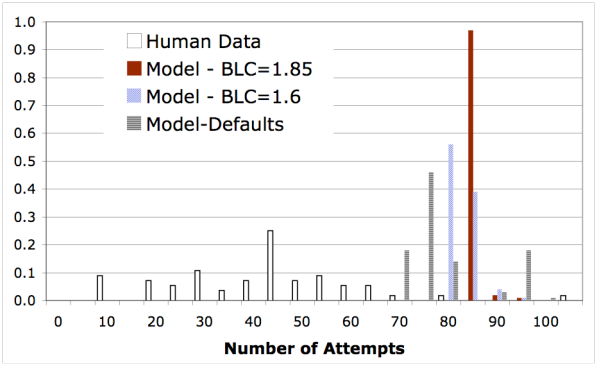

Figure 4. Distribution of attempts (errors not shown) for the 7's problems for the model with better settings and the human data distribution.

Discussion and Conclusion

The default settings for ACT-R lead the model's performance to match only part of the human data. Examining performance with a wider range of parameter settings suggests that individual differences are what give rise to the distribution that is observed. This is an interesting result, as it suggests ACT-R 6 produces peaked distributions of performance for each setting of parameters. This indicates that we may be able to fit to the average subject, but to fit the subjects' distribution we will have to use a set of parameter settings—the fit is not likely to be a single number, but will be matching of the distribution of individual differences.

Implications for Serial Subtraction

The analysis here confirm that utterance rate, noise, and base level activation are important in this task. In particular, the development of output mechanisms (speech rate) for architectures is important but somewhat unexplored.

There are a few further measures that would be useful for characterizing behavior on serial subtraction. For example, it would be interesting to know how the pace of subtractions, not just errors, changes over time. Do subjects get faster or slower over time? The error rate could increase because they are performing more subtractions, or it could be that they are performing them more poorly over time. Similarly, it would be interesting to know what errors subjects are making. Are they misretrieving the sub-answers, or are they forgetting to carry or to decrement? How does vocalizing while you are doing subtractions interfere with serial subtraction? We are working on these questions.

Implications for Architectures

This model and comparison show that distribution of response times and performance variables provide an additional useful, free, inexpensive, and strong constraint—on individuals and on population predictions.

The model was designed to exploit the integrated cognitive systems approach that lies at the core of ACT-R 6. The model performing the whole task including speaking illustrates this theoretical stance that is an important topic in current cognitive architecture research (Gray, 2007). Finally, we are also closer to a position to apply a set of theories of stress implemented as overlays to ACT-R (Ritter et al., 2007) to a sample data set to test the theories of stress on a task with detailed human data.

Acknowledgements

This project was supported by ONR (N000140310248, FER & LCK), the National Science Foundation (SBR 9905157, LCK), and from the Penn State University College of Health and Human Development (223 15 3605; LCK). The services provided by the GCRC of The Pennsylvania State University are appreciated (NIH Grant M01 RR 10732). We appreciate the dedicated assistance of E. Corwin and M. Stine in completing this project, as well as the research assistants in the Biobehavioral Health Studies Laboratory for subject recruitment, data entry, and data cleaning. Mark Cohen and three anonymous reviewers provided good comments.

References

Anderson, J. R. (in press). *How can the human mind exist in the physical universe?* New York, NY: OUP.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036-1060.

Gray, W. D. (Ed.). (2007). *Integrated models of cognitive systems*. New York: OUP.

Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The Trier Social Stress Test—A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology, 28*, 76-81.

Kudlielka, B. M., Buske-Kirschbaum, A., Hellhammer, O. H., & Kirschbaum, C. (2004). HPA axis responses to laboratory psychosocial stress in healthy elderly adults, younger adults, and children: Impact of age and gender. *Psychoneuroendocrinology, 29*, 83-98.

Nater, U. M., La Marca, R., Florin, L., Moses, A., Langhans, W., Koller, M. M., & Ehlert, U. (2006). Stress-induced changes in human salivary alpha-amylase activity—associations with adrenergic activity. *Psychoneuroendocrinology, 31*(1), 49-58.

Ritter, F. E., Bennett, J., & Klein, L. C. (2006). *Serial subtraction performance in the cycling study* (Tech. Report No. 2006-1): ACS Lab, College of IST, Penn State.

Ritter, F. E., Reifers, A., Klein, L. C., Quigley, K., & Schoelles, M. (2004). Using cognitive modeling to study behavior moderators: Pre-task appraisal and anxiety. In *Proceedings of the Human Factors and Ergonomics Society,* 2121-2125. Santa Monica, CA: HFES.

Ritter, F. E., Reifers, A. L., Schoelles, M., & Klein, L. C. (2007). Lessons from defining theories of stress for architectures. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 254-262). New York, NY: OUP.

Taylor, S. E., Gonzaga, G. C., Klein, L. C., Hu, P., Greendale, G. A., & Seeman, T. E. (2006). Relation of oxytocin to psychological stress responses and hypothalamic-pituitary-adrenocortical axis activity in older women. *Psychosomatic Medicine, 68*, 238-245.

Tomaka, J., Blascovich, J., Kelsey, R. M., & Leitten, C. L. (1993). Subjective, physiological, and behavioral effects of threat and challenge appraisal. *Journal of Personality and Social Psychology, 65*(2), 248-260.