Assessing Young Children's Hierarchical Action Segmentation

Meredith Meyer[1] (mermeyer@umich.edu),
Dare A. Baldwin[2] (baldwin@uoregon.edu), and Kara Sage[2] (kara@uoregon.edu)

[1]Department of Psychology, University of Michigan, Ann Arbor, MI 48103 USA
[2]Department of Psychology, University of Oregon, Eugene, OR 97403 USA

Abstract

Recognizing where action units begin and end is an early-developing skill that supports inferences about goals motivating others' action. One notable feature of goal-directed action is that segments are organized hierarchically. That is, action is interpreted as structured with respect to the goals and sub-goals of an actor, which can be recognized as corresponding to coarser- and finer-grained action units respectively. We report on the success of adapting a nonverbal paradigm to index hierarchical action segmentation in a developmental population. Results indicated that 3- and 4-year-old children, similar to adults in past studies, responded to segment boundaries with surges in attention that varied according to event granularity (e.g., fine- vs. coarse-grained). This effect was seen most strongly in children displaying superior memory for the events.

Keywords: action segmentation; event processing; development

Introduction

As social beings, we are routinely called upon to draw inferences regarding other people's goals and intentions based on observeable action. One initial step that aids in drawing such inferences is recognizing where action units begin and end within a stream of physically continuous motion; in other words, we can perceive a relatively continuous action stream as discrete segments, which we can map onto the internal and unobserveable goals of actors. For instance, while observing a person during meal preparation, we might segment and identify individual units of action such as cutting a vegetable, opening a microwave, or washing a dish. Studies of action perception indicate that people are quite consistent in how they segment observed action; people mark boundaries at roughly the same points within the motion stream, with units typically corresponding to what they perceive as initiation or completion of goals (Baldwin & Baird, 1999; Newtson, Engquist, & Bois, 1977; Zacks, Tversky, & Iyer, 2001). Action segmentation typically is subjectively experienced as effortless, generally proceeding automatically as part of our ongoing perception of human action (Hard, 2006; Zacks & Swallow, 2007).

People thus appear to be quite expert at segmenting continuous action into units. The apparent ease with which segmentation takes place is notable given the richness and complexity of the action stimulus itself. Human action is highly variable, evanescent, and lacks systematic pauses that reliably indicate where action units begin and end. Further, the underlying goal structure that motivates action is similarly rich and complicated, typically characterized by a structure corresponding to multiple and hierarchically-organized goals (e.g., Schank & Abelson, 1977; Zacks, Tversky, & Iyer, 2001).

Studies of action segmentation using both behavioral and neural measures have revealed that human observers perceive action in line with these hierarchical structures. For instance, people are capable of segmenting an action stream on multiple levels, ranging from coarse to fine (e.g., noting event boundaries of coarse-grained actions like “chop vegetable” at the onset and offset of the entire chopping event, or of finer-grained subunits at the onset and offset of each vertical movement of the knife). As in tasks assessing segmentation in general, tasks assessing hierarchical segmentation have also observed a high degree of consistency among people's judgments of where coarse and fine boundaries exist (e.g., Hard, 2006; Zacks et al., 2001a). Fine-grained judgments also align with coarse-grained judgments at rates higher than that expected by chance (Zacks, Tverky, & Iyer, 2001) and also typically are judged to occur at moments just preceding coarse-grained judgments, reflecting the presence of nested, or subordinate, units within the larger segmental structure (Hard & Tversky, 2011). Finally, fMRI studies suggest that activation levels in frontal and posterior areas vary depending on whether fine or coarse unit boundaries are observed, suggesting that hierarchical representation of action is psychologically real on a neural level (e.g., Zacks et al., 2001b).

The majority of behavioral research on action segmentation has relied on participants' explicit judgments of event boundaries, including the work outlined above. A necessary component of this work involves instructing the participants to note segments (e.g., with a key press), and it further requires clarification regarding the definition of "fine" and "coarse" (or equivalent terms) when investigating hierarchical processing. Although this work has produced compelling results supporting the presence of an automatic and hierarchical segmentation mechanism, the heavily verbal and explicit nature of the tasks is not well-suited to work with developmental populations, the population of interest in the current study.

Investigations into the development of segmentation are important for several reasons. First, it seems self-evident that sophisticated top-down mechanisms are at play when

we make inferences about the goals of others. Familiarity with others' actions and a realization that unseen goal states motivate action allow us to understand and make predictions about the actions of others. Infants and children, however, likely do not have such rich understanding of mental states, inviting the important question of how segmentation works in the absence of, as well as during the acquisition of, such knowledge. Does hierarchical action segmentation develop only after the aquisition of adult-like explicit understanding of action and goals, or might it exist as an early-developing perceptual processing style independent of explicit goal-state knowledge?

A Developmentally-Appropriate Methodology for Investigating Hierarchical Segmentation

Nonverbal looking time methodologies commonly used in infancy research have provided some promising means of investigating the developmental trajectory of segmentation. For example, Baldwin and colleagues used a familiarization method to examine ten-month-old infants' action processing. Infants who had first been familiarized to a simple action stream (consisting of a woman dropping a towel and bending down to pick it up) responded with increased looking time when pauses were inserted within action units (e.g., in the middle of bending down) as opposed to when pauses fell at action boundaries (e.g., at the moment the towel was grasped) (Baldwin et al., 2001). In another study by Saylor and colleagues, infants as young as nine months displayed a preference for dynamic human action that was accompanied by tones that matched action boundaries as opposed to action for which tones did not coincide with boundaries (Saylor et al., 2007).

Hespos and colleagues have also shown that even younger infants can detect action units presented within a sequence of continuous action. After habituating to a ball moving in two separate actions (e.g., Action 1 = ball placed *in* box, Action 2 = ball moved *over* bridge), 6- and 8-month old infants watched test sequences that either featured the two familiar actions within a stream of action (e.g., *in/behind/over*) or an entirely novel sequence (*on/behind/under*). Infants at both ages preferred to watch the novel sequence, suggesting that they recognized the units of action they had previously seen in isolation. In a second experiment, the authors also found that when infants *first* watched a stream of action in which the target action occurred, they similarly discriminated the target action in comparison to a novel action when these actions were presented in isolation during test (Hespos, Saylor, & Grossman, 2009).

The foregoing developmental studies all focused on preverbal infants, making use of standard familiarization, looking preference, or habituation/dishabituation methods. These looking time studies were directed at determining which of two events were preferred or yielded different attentional responses (i.e., a unit-completing pause vs. unit-interrupting pause, boundary-consistent tones vs. boundary-inconsistent tones, and familiar action vs. unfamiliar action). Although useful for addressing these comparisons, there are two disadvantages to standard looking time methods when the aim is to investigate hierarchical processing, the topic under consideration in the current study. First, typical investigations of action hierarchy compare perceptual responses among at least three levels of the action stream, e.g., within-unit, fine boundary, and coarse boundary. A methodology sensitive to a nested structure is therefore preferred, and the binary nature of standard looking time methods consequently is not well suited to this type of analysis. Second, the methodology is not amenable to investigations in older developmental populations, as the above looking time methods are rarely used beyond infancy.

Fortunately, recent work by Hard and colleagues (e.g., Hard, 2006; Hard & Recchia, 2006; Hard & Tversky, 2011) introduces a new method of examining the cognitive processes underlying segmentation that is both nonverbal and sensitive to processing of hierarchical structure, making it ideal for adaptation for the age used in the current study, namely preschool-aged children. (It is also likely adaptable downward to infancy, a topic we return to in the Discussion.) As this methodology forms the basis for the current investigation, a detailed description of its use and theoretical implications is in order.

Hard and colleagues were inspired by established paradigms used to examine hierarchical processing of text. In one such illustrative text processing study, participants saw one word at a time from a passage of text and advanced themselves word by word at their own pace by pressing a button. The length of time between button presses was recorded in this "moving window" method, with the expectation that increased cognitive load associated with processing demands would lead to longer delays between button presses. In particular, researchers found delays associated with the process of integrating past elements (words and/or phrases) into larger units. More specifically, participants typically spent longer periods of time on words located at the ends of unit boundaries. Further, this so-called "wrap up" effect was modulated by the level of the unit; reading times were longer for words located at the ends of clauses and even longer for sentence-final words (Haberlandt & Graesser, 1989).

To adapt this technique to study hierarchical action processing, Hard and colleague presented participants with a sequence of still-frame images sampled from regular time intervals from a movie of scripted dynamic human action (e.g., one still-frame image sampled every second). Participants advanced through these images with a button press, and the time between presses was recorded. Following this "slideshow", participants saw the live action footage from which the still images had been sampled and marked with a button press the locations of action boundaries (hereafter, 'breakpoints'). Participants completed this explicit segmentation task a total of three times, providing separate judgments on fine, intermediate, and coarse levels.

Results from the slideshow task established that participants spent a longer time looking at images close in time to moments they judged to be breakpoints, in comparison to images taken from within action units. Further, similar to the results obtained in text processing, this effect was modulated by the level of the breakpoint, with slides close in time to moments judged as coarse-grained breakpoints receiving the most looking time and those near fine-grained breakpoints receiving the least. These results, collectively dubbed the dwell time effect, provided evidence that hierarchical segmentation occurs as part of real-time perception, without depending on processes associated with explicit segmentation judgments. (That is, the modulation based on the hierarchical status of an event unit occurred during participants' watching of the slideshow; since participants did not make explicit segmentation judgments until later, one can conclude that there are cognitive signatures of hierarchical processing that can be detected independent of what results from an explicit intention to segment.)

Hard and colleagues explained their results by suggesting that breakpoints are cognitively privileged, demanding additional attention and processing in order for observers to consolidate and integrate action units into a hierarchical action representation. Interestingly, these authors additionally found that participants' later explicit memory for the action sequences predicted higher degrees of modulated dwell times; the more events participants recalled from the sequences (tested after both the slideshow and explicit segmentation phases), the more their dwell times reflected the hierarchical modulation effect. Thus, it appears that the degree to which action is successfully encoded and retrieved relates to the way it is processed during observation.

The results obtained by Hard and colleagues, as well as a later replication by Meyer and colleagues (Meyer et al., 2010) are also consistent with Event Segmentation Theory, an account of action segmentation developed by Zacks and colleagues (e.g., Kurby & Zacks, 2007; Zacks et al., 2007). According to this theory, segmentation is a consequence of prediction generation, a spontaneous, online process that integrates incoming sensory information with prior knowledge and learning. Event segments correspond to periods in which prediction error rate is low; the observed action is consistent with the system's predictions. For example, within the event of chopping a vegetable, the system generates accurate predictions of further chopping based the person's movements as well as prior knowledge about vegetable preparation. Segment boundaries, in contrast, arise when prediction error rate is high; to extend the example above, such boundary moments are likely to occur at the completion of a segment (e.g., finishing chopping) and before the onset of another segment (e.g., opening the microwave door), because these moments are associated with reduced ability to predict the content of the second event. In order to update the system at moments of reduced predictability, observers are believed to automatically increase attention to the perceptual attributes of the action stream. The idea that transient surges in attention are required at boundaries is consistent with Hard and colleagues' findings; further, dwell time findings also suggest that the surges are affected by the granularity of the events being witnessed, with coarser-grained unit boundaries requiring the most attention (and likely related to the highest degree of prediction error), and finer-grained unit boundaries eliciting less (and likely related to relatively lower prediction error).

Overview of the Current Study

Hard and colleagues' work thus demonstrated that breakpoints are processed differently than within-unit moments, with the detection of boundaries resulting in a transient increase in cognitive processing load that varies depending on the granularity of the segment. Better memory for events was also related to the dwell time effect. In the current study, we adapted this method for use with preschool-aged children with only two major changes. First, rather than match dwell times on the slideshow portion to participants' own explicit judgments regarding breakpoints, we *a priori* noted the location of breakpoints within a stream of action and grouped children's dwell times according to these experimenter-determined judgments. This change was necessary because of the concern described above that young children could not manage an explicit segmentation task that demanded they recognize the difference between coarse- and fine-level units. Second, we also included a measure of participants' memory, but instead of asking children to recount as many events as they could remember (the method used to assess memory in the adult work), we used a forced choice memory task to assess children's memory. Again, this change was instituted to make the procedure easier and manageable by a younger population.

Three less significant changes were also instituted in our adaptation of the methodology. Namely, first, we had a brief training period during which we taught children to click a mouse in order to advance through the slideshow. Second, we used child-friendly action depicting someone assembling toys rather than the more complicated action sequences chosen for studies with adults (e.g., assembling furniture, cleaning a room, etc.). Finally, we scripted a somewhat simpler action sequence designed to feature three levels of action (within-unit, fine, and coarse), rather than the four levels used in past studies of dwell time (i.e., within-unit, fine, intermediate, and coarse).

We predicted that children, like adults, would show an increase in dwell time for breakpoints in comparison to non-breakpoints, and further that this would be modulated by the level of the breakpoint, with coarse-level breakpoints receiving the most dwell time and fine-level breakpoints receiving relatively less dwell time. We also predicted that children's memory for events would relate to this effect, with modulation seen more strongly in children with better memory.

Method

Stimuli

Images for the main slideshow viewing task were created by extracting one image every second from an 88-second movie clip depicting an individual interacting with three toys, one toy at a time. The individual first briefly smiled and waved while looking into the camera, then assembled a stack of plastic rings, next nested a series of cups, next placed two stuffed animals into a box, and finally briefly waved again. Images from this sequence were classified as depicting greeting or ending phases (waving portions), within-unit action, fine-unit breakpoints, or coarse-unit breakpoints. Examples of a within-unit, fine-unit, and coarse-unit breakpoint image are depicted in Figure 1. ("Waving" images were not of theoretical interest and were only used to engage children.)

We also selected images for a first training phase consisting of ten child-friendly pictures (e.g., Elmo, a kitten, a puppy, etc.) and images for a second training phase consisting of approximately 30 images regularly sampled (one every second) from a brief movie of a woman potting a plant.

Participants and Procedure

Participants were 14 3-year-olds (M = 41.5 months, SD = 3.96; 9 male) and 12 4-year-olds (M = 53.08, SD = 3.70; 5 male). The experiment consisted of two brief training phases, followed by the main slideshow task and the memory test. Images in the slideshow were presented on a Macintosh G4 computer with a 19.5" x 12" monitor, and children sat approximately three feet away. Children sat alone or on a parent's lap; if they were on the parent's lap, the parent was asked to wear a visor and avoid looking at the monitor.

Within-unit slide depicting actor in the middle of placing medium ring

Fine-unit slide depicting actor completing placement of medium ring

Coarse-unit slide depicting completed ring assembly; actor next moves to cups

Figure 1: Within-unit, fine-grained, and coarse-grained slides from the main slideshow

Children started with the first training phase. The experimenter "clicked through" the first three images to demonstrate and then instructed the child to click. The experimenter then prompted the child to click through the second training phase. Finally, children clicked through the main slideshow. During this last session, children's dwell times for each slide were coded by a trained coder out of sight from the child using Psychtoolbox (Brainard, 1997).

The memory task consisted of eight forced-choice recognition and recall items. Three questions asked children to select which of two toys the experimenter had played with, three questions asked children to select which of two actions the experimenter had performed, and the last two questions probed children's memory for temporal order of events.

Results

Calculating Dwell Time Scores

Outlying looking times (>2 standard deviations above the group mean) were removed. Data were then subjected to same treatment used in Hard & Tversky (2011) and Meyer et al. (2010), namely 1) log-transforming data to remove positive skew, 2) calculating residuals off power functions fitted individually to participants' looking times, and 3) creating dwell time scores by dividing mean looking times per slide type by the standard deviation of times within those types. Here, we grouped slides into three groups according to whether they appeared as a) coarse-unit or immediately before or after the slide designated as coarse-unit, b) fine-unit or immediately before or after the slide designated as fine-unit, or c) within-unit. (Hereafter these classifications are referred to simply as coarse-grained, fine-grained, or within-unit slides. This classification was used first because there were not enough coarse-grained slides to yield stable mean measures of looking times, and second because we expected children's looking behavior to be less organized than that of adults, with less coordination between perception and the motor response of clicking the mouse.) The first step of log-transforming data is standard in looking time analyses; the second step of calculating residuals was used because of viewers' tendency to look for a long time at the initial few images and then to increase in advancement rate after this initial phase; and finally, the third step was used to correct for the fact that means for breakpoints were obtained from fewer data points than means for within-unit slides (using means divided by standard deviations essentially provides a measure of effect size). Importantly, the second step, namely entering residuals into the calculation of the dwell time score, creates the possibility of negative data points (i.e., observed data lying under the

predicted power function yielding negative values); however, it should be noted that lower dwell time scores nevertheless still indicate lower looking, and higher dwell time scores indicate more looking.

Dwell Time Score Analysis

A 2 (age: 3-year-old vs. 4-year-old) x 3 (level: within-unit, fine-unit, coarse-unit) mixed between-within ANOVA was run on dwell time scores, with age as the between-subjects variable and level as the within-subjects variable. Level was marginally significant, $F(1.52, 36.35) = 3.02$, $p = .07$, and, as predicted, characterized by a significant linear trend, $F(1, 24) = 4.91$, $p = .04$ (M_{within}= -.04, SEM = .04; M_{fine} = .07, SEM = .03; M_{coarse}=.17, SEM = .09). Age group was not significant, $F(1,24) = .61$, $p > .05$; nor was the age group x level interaction, $F(1.52, 36.35) = 1.72$, $p > .05$. (Greenhouse-Geisser adjusted *df* reported when appropriate due to violations in sphericity.)

To explore the possibility that memory was related to the dwell time modulation, we also ran two separate analyses examining dwell time scores in individuals whose memory scores were above the median score of 7.5 (n = 13) vs. below (n = 13). Here, a one-way ANOVA examining dwell time scores across the three different levels yielded significant effects only in the high-memory group, $F(2, 24) = 3.56$, $p = .04$, with the predicted significant linear trend, $F(1, 12) = 5.4$, $p = .04$ (M_{within} = -.05, SEM = .06; M_{fine} = .04, SEM = .04; M_{coarse} = .3, SEM = .09). The same one-way ANOVA was not significant for the low-memory group, $F(2,24)=.49$, $p > .05$ (nor were linear or quadratic trends) (Figure 2).

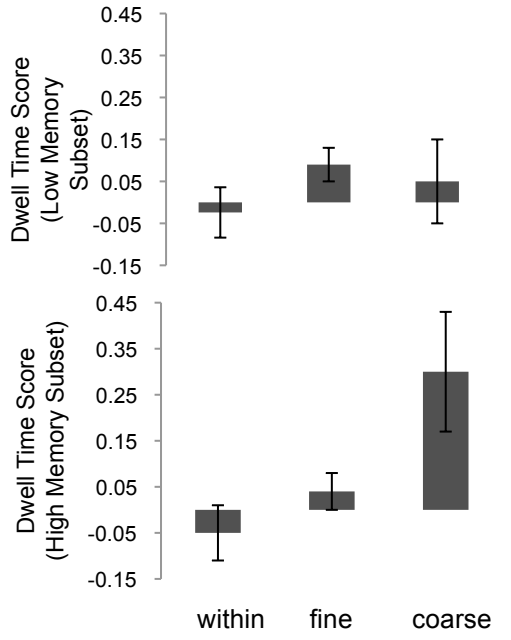

Figure 2: Dwell time scores to within-unit, fine-grained, and coarse-grained image classes in individuals with low memory scores (above) and high memory scores (below). Only high memory individuals showed the predicted linear trend, $p < .05$.

Discussion

In summary, 3- and 4-year-old participants showed a linear trend in their dwell times whereby dwell times were longest for coarse-grained images and shortest for within-unit images, paralleling findings from studies of adults by Hard and colleagues (Hard & Tversky, 2011). Further, hierarchy-related dwell time modulation in our participants was only strongly observed among individuals who had scored high on the memory task, a finding that is reminiscent of Hard and colleagues' discovery that adults' memory recall related to strength of dwell time modulation obtained in their study. Our findings thus suggest three important points: First, dwell time modulation is a robust and valid phenomenon even within a developmental population; second, use of the dwell time paradigm is capable of providing another window into the cognitive processes underlying segmentation even within child participants; and third, children's memory for events appears to be related to the dwell time phenomenon.

The fact that children's memory appeared to matter for dwell time warrants more in-depth investigation. It is likely that our memory test was too easy for most of our participants; indeed, children scoring above the median score of 7.5 were in fact children who received perfect scores on the measure. Developing a test that yields more variation in scores is one important pursuit for the future. Further, our results are not at all demonstrative of the causal role of hierarchical segmentation in memory for action. A number of associated abilities could have contributed to children's performance on the memory task, including transient mood or attentional states, executive function, or engagement with the task; further, these same factors may also have contributed to children's behavior on the slideshow task as well.. In any event, the fact that memory does at the very least relate to the dwell time effects that we observed invites further investigation into the phenomenon.

Our findings also open up a number of broader questions suitable for future investigation. One question that arises is the degree to which dwell time is dependent on processes related to explicit understanding of goal states. Although we chose 3- and 4-year-olds as a population that may not have entirely adult-like mental state and goal understanding, it is likely that they possessed at least some understanding of the actions witnessed in our movie (i.e., stacking plastic rings, nesting cups, and putting things into boxes). In particular, investigating issues of top-down knowledge acquisition and its role in contributing to dwell time effects is interesting in light of Event Segmentation Theory, which holds that prediction is the central process involved in perceiving action segments. To what degree is *explicit* prediction related to this process? Would the same results obtain if we showed children movies of less familiar actions in which it would be harder to predict each next step of the actor? Would theory of mind or other standardized tests of mental state understanding relate to the dwell time effect?

Our findings also offer an exciting direction for future investigations within infants. As described above, use of

standard looking time paradigms has revealed clear evidence for infants as young as nine months being able to segment an action stream, a notable finding in light of infants' relatively impoverished understanding of goals and intentions (e.g., Baldwin et al., 2001; Saylor et al., 2007). Although these studies represent a compelling demonstration of infants' action processing skill, the adaptation of dwell time methodology to infants has the potential to further expand our understanding of the developmental trajectory underlying the segmentation process, particularly with respect to hiearchical processing. We are currently developing a methodology in which infants' motor movements (namely, patting a touchscreen) result in advancement of slides.

As they stand now, however, our results are still cool for the following couple of reasons.

Concluding paragraph.

Acknowledgements

This research was supported by the U.S. Office of Naval Research, award no. N000140910187 to the second author. We thank Jeffrey Loucks for assisting in MatLab programming.

References

Baldwin, D., & Baird, J. A. (1999). Action analysis: A gateway to intentional inference. In P. Rochat (Ed.), *Early social cognition*, (pp. 215–240). Hillsdale, NJ: Lawrence Erlbaum Associates.

Baldwin, D., Baird, J., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development, 72*, 708–718.

Haberlandt, K., & Graesser, A. C. (1989). Processing of new arguments at clause boundaries. *Memory & Cognition, 17*, 186-193.

Hard, B. (2006). Reading the language of action: Hierarchical encoding of observed behavior. Doctoral dissertation, Stanford University.

Hard, B., & Recchia, G. (2006). Reading the language of action. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, (pp. 1433-1439), Vancouver, CA.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences, 12,* 72-79.

Newtson, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology, 35*, 847–862.

Saylor, M. M., Baldwin, D., Baird, J. A., & LaBounty, J. (2007). Infants' on-line segmentation of dynamic human action. *Journal of Cognition and Development, 8,* 113-128.

Schank, R. C., & Abelson, R. P. (1977). Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Hillsdale, NJ: Lawrence Erlbaum Associates.

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science, 28*, 979–1008.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., et al. (2001b). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience, 4*, 651–655.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind/brain perspective. *Psychological Bulletin, 133*, 273-293.

Zacks, J. M. & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science, 16,* 80-84.

Zacks, J. M., Tversky, B., & Iyer, G. (2001a). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General, 130*, 29–58.