Supplemental Material

How Does Aging Impact Decision Making?

The Contribution of Cognitive Decline and Strategic Compensation

Revealed in a Cognitive Architecture

Hanna B. Fechner, Thorsten Pachur, and Lael J. Schooler

Details on why Participants Were Excluded

In this section, we give an overview of the number of excluded participants and the underlying reasons: 24 were excluded because they could not successfully complete the learning phase or main phase of the experiment (17 in the take-the-best group of older adults, seven in the tallying group of older adults), seven because of a technical problem in recording the responses in one of the cognitive covariates, that is, the working memory test battery (two in the take-the-best group of older adults, two in the take-the-best group of younger adults, one in the tallying group of older adults, and two in the tallying group of younger adults), 12 because we had measured timing errors of more than 300 ms for the presentation of tones in more than 10 of their trials (seven in the take-the-best group of older adults, four in the tallying group of older adults, and one in the tallying group of younger adults), four because of crashes of the experimental software triggered by accidentally pressing one of the Windows keys (one in the take-the-best group of older adults, two in the take-the-best group of younger adults, and one in the tallying group of younger adults), and three because a different version of the experiment was run (one in the take-the-best group of older adults, one in the take-the-best group of younger adults, and one in the tallying group of older adults).

Learning Phase of the Empirical Study

The learning phase consisted of four tasks designed to familiarize participants with the materials and tasks of the study. In a *tone-familiarization task*, participants first adjusted the volume with the assistance of the experimenter. Then they heard birdsongs in a random sequence and indicated which species was singing until they had responded correctly five times in a row for all three tones (i.e., 15 identifications). The birdsongs can be downloaded from http://animaldiversity.ummz.umich.edu They were recorded by Douglas von Gausig, Naturesongs.com, P.O. Box 490, Clarkdale, AZ 86324, and are available under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported License, http://creativecommons.org/licenses/by-nc-sa/3.0/

In an *attribute-familiarization task*, participants were first informed about the five attributes of mammals (see Materials of the empirical study) and learned that high values often indicated a long life span. Then participants saw the attributes in a random sequence and indicated which of the two attribute symbols (randomly allocated to the left and right side of the display) suggested a long life span until they had responded correctly five times in a row for all five attributes (i.e., 25 identifications). Both familiarization tasks were repeated once.

In a *tone-counting-practice task*, on each trial participants were told to listen to or count one or two kinds of birdsongs. They then heard a random sequence of seven, eight, or nine birdsongs and typed their responses. The other settings of the practice corresponded to those of the tone-counting task in the main phase of the empirical study. The practice continued until participants had listened in the no-counting condition and responded correctly in the two counting conditions four times in a row (i.e., 12 runs). Then the practice of the tone-counting task was repeated.

In a *decision-practice task*, participants learned about the importance of the attributes for making the decisions (see Materials of the empirical study) and were instructed to use take-the-best or tallying. Participants using take-the-best should compare the two animals in terms of their attribute values, starting with the most important attribute. In the case of different attribute values, the animal with the high value should be chosen as the one with the longer life span; in the case of no difference in the attribute values, participants should proceed to the second-most important attribute, and so forth (two examples were given, one for a situation where the first attribute discriminated and one where the second attribute discriminated). Participants using tallying should sum the information for each animal separately, by adding 1 for a high value and adding nothing for a low value, and then compare the sums for both animals; the animal with the higher sum should be chosen as the one with the longer life span (an example was given where the attribute values of one animal summed to 3, while the attribute values for the other summed to only 2).

Then participants practiced the decision task using their instructed strategy. On each trial, participants saw the attribute information of two animals and indicated which animal would have the longer life span (see details in Figure 1) until they had chosen the animal predicted by their instructed strategy 10 times in a row (i.e., 10 correct decisions; guessing trials did not count); otherwise the practice block was repeated (with a different random order of trials and random screen positions for attributes). Then a second decision-practice block was presented. Each of these practice blocks consisted of 11 paired comparisons, with two pairs of each item type, and one pair that required guessing; for four of these pairs (i.e., two pairs of item type 1 and 2) take-the-best predicted a different choice from that of several integrative decision strategies (among them tallying). These pairs differed from the pairs of the main phase and the other practice block for each participant.

Cognitive Measures for Younger and Older Adults

For characterizing the participant samples we used several cognitive measures. At the beginning of each session we applied the Digit Symbol Substitution Test from the Wechsler Adult Intelligence Scale (von Aster, Neubauer, & Horn, 2009) as a measure of processing speed and a proxy for fluid intelligence (cf. Lindenberger, Mayr, & Kliegl, 1993). After the main experiment, participants filled out a questionnaire on the tasks and their strategies. Then they worked on two tasks from a working memory test battery, that is, a memory-updating task and an operation-span task (Lewandowsky, Oberauer, Yang, & Ecker, 2010; using the

Psychophysics (PST) Toolbox Version 2.54 for MATLAB; Brainard, 1997; Pelli, 1997). Subsequently, a vocabulary test was used (Lehrl, Merz, Burkard, & Fischer, 1991) as a measure of verbal knowledge and a proxy for crystalline intelligence (cf. Spot-a-Word; Lindenberger et al., 1993). Finally, participants filled out questionnaires on handedness (Oldfield, 1971) and sociodemographic variables.

Working memory test battery: Memory updating. Participants encoded an initial set of digits (set size: three to five), that were presented in separate frames on the screen, updated these digits by applying (two to six) displayed arithmetic operations, and recalled the resulting digits at the end of each trial; 15 trials were presented that were constructed by crossing the set sizes with the number of updating operations. For participants in the take-the-best condition accuracy for memory updating was higher in the younger group (M = .58, SD = .19) than in the older group (M = .39, SD = .16), $t(70)$ = 4.686, p < .001. Similarly, for participants in the tallying condition accuracy for memory updating was higher in the younger group (M = .57, SD = .20) than in the older group (M = .41, SD = .16), $t(70)$ = 3.713, p < .001.

Working memory test battery: Operation span. Participants were presented with an alternating sequence of arithmetic equations, which they had to judge as correct or incorrect, and consonants (list length: four to eight), which they had to remember and recall serially at the end of each trial; 15 trials with 3 trials per list length were presented.

For participants in the take-the-best condition accuracy in the memorization task of the operation span was higher in the younger group (M = .77, SD = .15) than in the older group (M = .56, SD = .20), $t(70)$ = 4.996, p < .001. Similarly, for participants in the tallying condition accuracy in the memorization task of the operation span was higher in the younger group (M = .75, SD = .18) than in the older group (M = .57, SD = .21), $t(70)$ = 3.661, p < .001.

For participants in the take-the-best condition accuracy in the processing task of the operation span was higher in the younger group (M = .93, SD = .08) than in the older group (M = .87, SD = .08), $t(70)$ = 3.126, p = .003. However, for participants in the tallying condition accuracy in the processing task of the operation span in the younger group (M = .90, SD = .05) was comparable to the accuracy in the older group (M = .85, SD = .16), $t(70)$ = 1.681, p = .097.

Digit symbol substitution. Participants saw a table that related nine digits to different symbols. They were given a larger table that held only digits for which they had to fill in as many symbols as possible within 120 s. For participants in the take-the-best condition the number of correct substitutions in the younger group (M = 88.14, SD = 13.88) was higher than the number in the older group (M = 65.06, SD = 14.45), $t(70)$ = 6.915, p < .001. Similarly, for participants in the tallying condition the number of correct substitutions in the younger group (M = 85.61, SD = 12.51) was higher than the number in the older group (M = 67.25, SD = 11.83), $t(70)$ = 6.400, p < .001.

Vocabulary. Participants saw 37 rows of five letter strings each (consisting of one word and four pronounceable nonwords) and had to determine which of the five strings would be a word that they would know. For participants in the take-the-best condition the number of correct word identifications in the older group (M = 32.86, SD = 1.53) was higher than the number in the younger group (M = 31.03, SD = 1.96), $t(70)$ = -4.415, p < .001. Similarly, for participants in the tallying condition the number of correct word identifications in the older group (M = 33.47, SD = 1.66) was higher than the number in the younger group (M = 30.83, SD = 3.35), $t(70)$ = -4.232, p < .001.

Handedness. Participants indicated their preferred hand for using 10 everyday devices. For participants in the take-the-best condition the laterality quotient (ranging from -100, indicating left-handedness, to 100, indicating right-handedness) in the older group (M = 77.54, SD = 39.44) did not differ from the quotient in the younger group (M = 64.42, SD = 45.64), $t(70)$ = -1.305, p = .196. However, for participants in the tallying condition the laterality quotient in the older group (M = 77.67, SD = 34.89) tended more toward right-handedness than the quotient in the younger group (M = 49.25, SD = 62.65), $t(70)$ = -2.377, p

= .020.

Comparison of Observed Data and Simulation Results for Younger Adults

In this section, we describe and compare the simulation results from the ACT-R models with the observed data for younger adults, who were instructed to apply either take-the-best or tallying. The simulations were based on default or typical parameter values for younger adults derived from the literature and on the strategy instructions given to participants. Both simulation results and observed data are shown in Figure S1. The models captured the main patterns of the data well, but produced also some noteworthy deviations.

Take-the-best. When examining the results for take-the-best, under working memory load (i.e., 1 and 2) and increasing search requirements, the ACT-R simulations underestimated the observed accuracy in strategy execution (Figure S1, top left; $RMSD$ = .05, r = .34). At the same time, while correctly predicting the main pattern of reaction times, under working memory load 1, the simulations overestimated the observed response times, and underestimated them under working memory load 2 (Figure S1, middle left; $RMSD$ = 1076.95, r = .92). In both cases, the simulations matched the pattern in accuracy in tone counting, but underestimated the observed accuracy level (Figure S1, bottom left; $RMSD$ = .05, r = .93).

In general, it is possible that some of the younger adults applied similar strategic shifts that reallocate resources toward decisions as described and tested for the data of older adults; this could increase execution accuracy and speed for decisions compared with performance with equal task priority as implemented in the model. However, given that the model with equal task priority underestimated the observed accuracy in tone counting, these could primarily be shifts that do not impair the accuracy in the tone-counting task (e.g., the monopolizing shift).

Under working memory load 1, a possible reason for the overestimation of response times and underestimation of accuracy for decisions and tone counting might be that participants relied on auxiliary strategies for tone counting. Such auxiliary strategies might be more applicable under low (compared with high) working memory load (e.g., counting with fingers instead of storing counts in memory, subvocalizing, creation of visual mental representations, as reported by some participants). Auxiliary processes may reduce response times and increase accuracy for decisions (because of fewer interruptions from tone counting followed by error-prone restoring of decision-relevant information from memory) and for tone-counting (because of less necessity of error-prone restoring of counts), compared with performance that fully relies on working memory functioning (including restoring intermediate results for both tasks) as implemented in the model.

Under working memory load 2, as for the underestimation of response times and accuracy in decisions and tone counting, it is possible that participants in this most difficult condition of the experiment still attempted to reach a high accuracy as they obtained their bonus only with an overall accuracy of at least 50 % in both tasks. This could have led them to repeat some processing steps of take-the-best, when having been interrupted by the tone-counting task; this could have increased response times but also accuracy in strategy execution (and probably also tone counting), compared with executing all steps of take-the-best only once as implemented in the model.

Tallying. As for the results of tallying, the ACT-R simulations matched the observed accuracy pattern for strategy execution, but slightly underestimated accuracy under working memory load 2 (Figure S1, top right; *RMSD* = .02, *r* = .77). The simulations quantitatively underestimated the observed response times, without showing the unexpected observed effect of item type (exactly as it should be when applying tallying; Figure S1, middle right; *RMSD* = 1,890.39, *r* = .79). In addition, the simulations overestimated the accuracy in tone counting (Figure S1, bottom right; *RMSD* = .06, *r* = .87).

A possible reason for the underestimation of response times might be that participants applied additional or alternative processes beyond the instructed tallying strategy: For instance, participants could have executed multiple loops over the same steps of tallying, which would increase their response times (and potentially also accuracy in strategy execution), compared with executing tallying once as implemented in the model. Furthermore, it is possible that some participants executed a full summation of attribute values only for the first alternative, but truncated the summation for the second alternative with a stop when the second alternative exceeded the first, which might also reduce response times. In addition, some processes may facilitate performance when attribute patterns are similar (i.e., for item types for which only less important attributes discriminate between alternatives), which could explain the unexpected effect of item type: For instance, it is possible that participants scanned the display to detect differences between alternatives and limited information integration to only differing attributes; this might decrease response times for item types with similar alternatives. As such alternative processes might require resources that are also used for tone counting, engaging in these processes might also have reduced accuracy in tone counting.

In sum, these deviations seem to suggest that some participants may have relied on auxiliary or alternative strategies. As the exact underlying processes are unknown and may differ for each participant, we opted not to include such additional processes in our models to account for these deviations. We also opted to keep ACT-R's parameters in each resource (for details, see Table 1) at their default or typical values for younger adults and not to adjust these parameters. Such adjustments would only allow us to scale the global performance level, but would not necessarily account for the observed mismatch in specific conditions. Instead, we treated the simulation results as the outcome of prototypical examples of the two decision strategies, executed without any further processes beyond the instructed strategies' algorithms and with equal priority to the decision and tone-counting task.

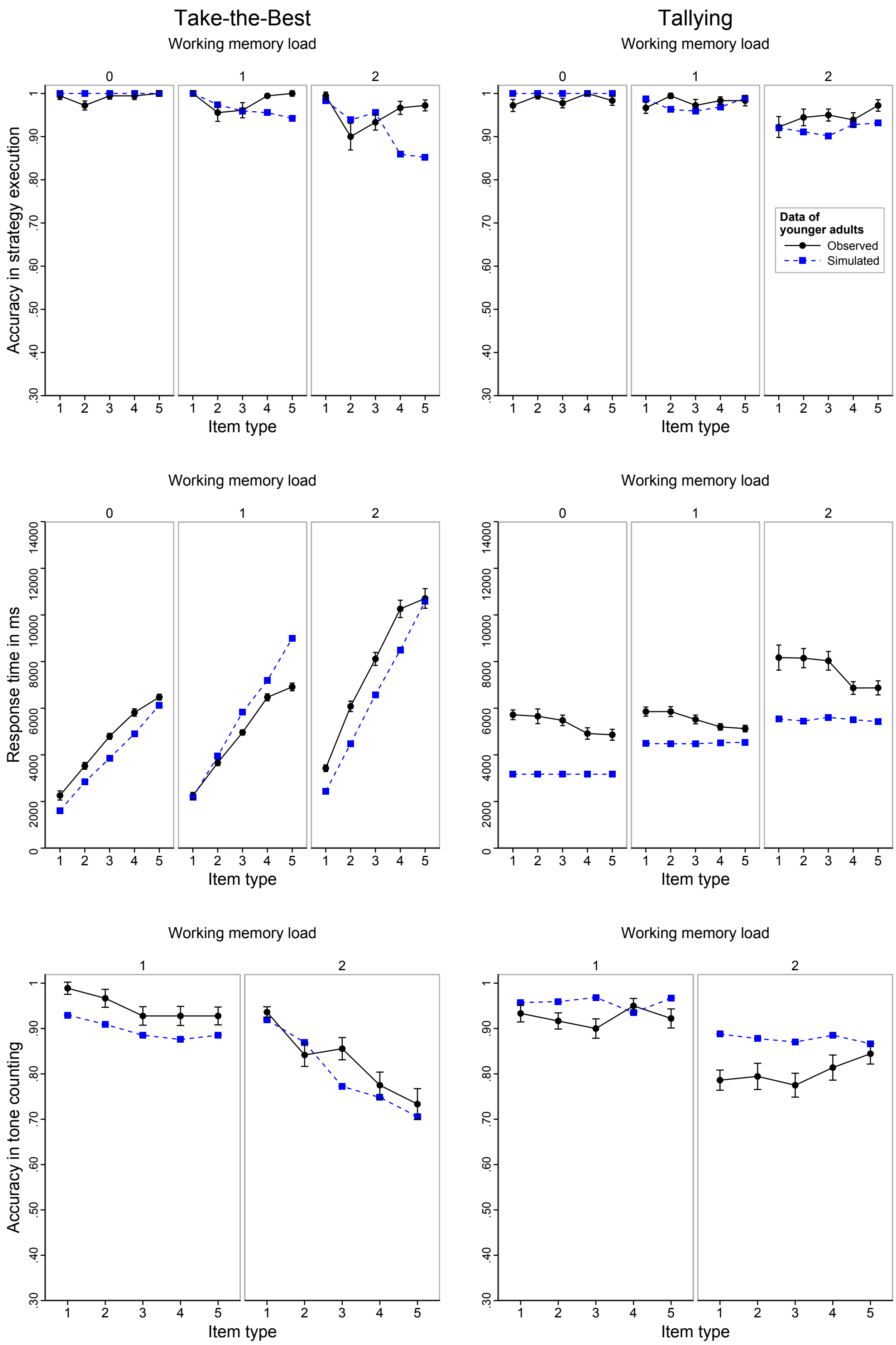

Figure S1. Results for younger adults using take-the-best (left column) and tallying (right column), for both the empirical study (black circles with solid lines) and simulations (blue [dark gray] squares with dashed lines). Displayed are the accuracy in strategy execution (top row), the response times (middle row), and accuracy in tone counting (bottom row) across the working memory load conditions (i.e., 0-2 for decisions, and 1-2 for tone counting) and item types that vary the search requirements for take-the-best (i.e., 1-5 attributes). Error bars represent standard errors corrected for within-subject designs.

Effects of the Strategic Shifts in the Tallying Condition

In this section, we present the effects of the three strategic shifts in the tallying condition that we had previously examined for the take-the-best condition. Tallying was explained to participants in its alternative-wise variant that involves examining the attribute values of the first and then of the second alternative. Consequently, the monopolizing shift (according to which decision strategies can monopolize resources for several processing steps) was implemented for tallying as follows: A representation of decision-relevant information occupies the imaginal buffer while each attribute value is examined and the corresponding score in the problem state is updated (in case of a high attribute value). If there are more attribute values to be examined, the imaginal buffer is freed for tone counting before proceeding to the next attribute value. The time-limit shift (including the threshold setting) and selective-aspects shift were implemented as described for take-the-best.

Accuracy in strategy execution (strategic compensation). For each shift, the top row of Figure S2 shows the simulated and observed accuracy in strategy execution of tallying. The monopolizing shift decreased accuracy under increasing working memory load (compared with a situation without the shift), as this shift interrupted the execution of tallying more often than necessary and thus created opportunities for errors (Figure S2, top left). The time-limit shift left accuracy unaffected (Figure S2, top central; for details, see next section on response times). The selective-aspects shift increased accuracy under high working memory load (Figure S2, top right), as with the shift fewer cognitive processes were executed for tone counting that could interrupt the decision. Although this shift may represent a means to counteract potential accuracy decline, it overestimated the observed accuracy of older adults. The remaining outcomes of the shifts were consistent with older adults' execution accuracy across conditions.

Response times (strategic compensation). The middle row of Figure S2 displays the simulated and observed response times for tallying. The monopolizing shift interrupted the execution of tallying more often than necessary and increased response times compared with a situation without the shift (Figure S2, middle left). This was the case without and with working memory load, as with the shift the imaginal buffer was cleared after processing each attribute value, even when the imaginal buffer was not required by the concurrent task. The time-limit shift (with the same threshold as for take-the-best) did not lead to a change in response times. This was the case because under high demands tallying's response times were substantially shorter than those of take-the-best (Figure S2, middle central) and were thus barely affected by the time limit for multitasking. The selective-aspects shift decreased response times for tallying under high working memory load (compared with a situation without the shift), as with the shift fewer cognitive processes were executed for tone counting (Figure S2, middle right). Taken together, for tallying only the selective-aspects shifts decreased response times under high demands and had therefore the potential to counteract decline, whereas the monopolizing shift increased the response times and the time-limit shift had no effect. It remains unclear whether older adults engaged in such shifts while applying tallying, as their data were captured by assuming cognitive decline.

Accuracy in tone counting (strategic compensation). The bottom row of Figure S2 shows the simulated and observed accuracy in tone counting when applying tallying. The monopolizing shift slightly decreased accuracy (compared with a situation without the shift). This is because the shift created more interference between tasks and prolonged response times, such that there were more opportunities to commit errors during tone counting (Figure S2, bottom left). With the time-limit shift, accuracy in tone counting was unaffected when executing tallying (see previous section; Figure S2, bottom middle). The selective-aspects shift decreased accuracy under high working memory load (compared with a situation without the shift), but overestimated older adults' observed decline (Figure S2, bottom right). In sum, the shifts did not lead to an increase in accuracy in tone counting. However, shifts that lower the accuracy for the benefit of decision performance would be consistent with older adults' observed accuracy in tone counting.

Overall, for tallying, strategic shifts were not necessary to explain the characteristics of older adults' observed performance (which was accounted for by assuming cognitive decline), but—with slight changes in their implementation—the shifts might bear the potential to counteract effects of cognitive decline in at least one of the two tasks (as was already the case for the selective-aspects shift).

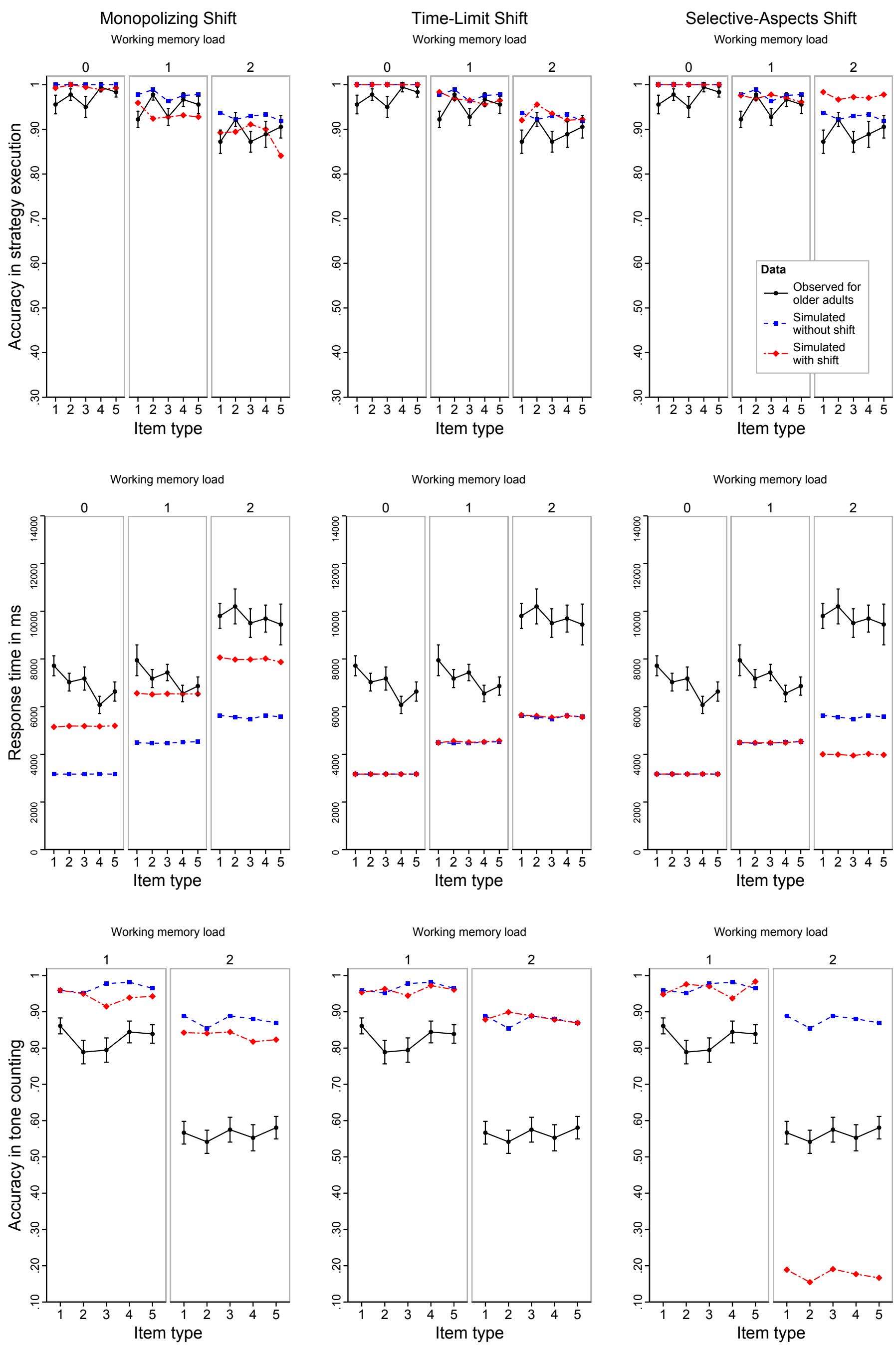

Figure S2. Simulation results of strategic shifts for the tallying condition. Displayed is the accuracy in strategy execution (top row), response times (middle row), and accuracy in tone counting (bottom row) for the monopolizing shift (left column), the time-limit shift (middle column), and the selective-aspects shift (right column) when applying tallying, across the working memory load conditions (i.e., 0-2 for decisions, and 1-2 for tone counting) and item types that vary the search requirements of take-the-best (i.e., 1-5 attributes). Simulation results for strategic compensation with each shift are displayed as red (medium gray) diamonds with dot-dashed lines; for comparison, results obtained without shifts are shown as blue (dark gray) squares with dashed lines. Simulations with and without shifts were run with the default parameter values of younger adults. The observed data of older adults are shown as black dots with solid lines; error bars represent standard errors corrected for within-subject designs.

References for Supplemental Material

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Lehrl, S., Merz, J., Burkard, G., & Fischer, B. (1991). *Mehrfachwahl-Wortschatz-Intelligenztest MWT-A (Parallelform zum MWT-B) [Multiple choice vocabulary intelligence test MWT-A (Parallel form for MWT-B)]*. Erlangen, Germany: Perimed Fachbuch-Verlagsgesellschaft mbH. plus Testblätter [Record forms] (2007). Balingen, Germany: Spitta.

Lewandowsky, S., Oberauer, K., Yang, L.-X., & Ecker, U. K. H. (2010). A working memory test battery for MATLAB. *Behavior Research Methods*, *42*, 571–585. doi:10.3758/BRM.42.2.571

Lindenberger, U., Mayr, U., & Kliegl, R. (1993). Speed and intelligence in old age. *Psychology and Aging*, *8*, 207–220. doi:10.1037/0882-7974.8.2.207

Oldfield, R. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*, 97–113. doi:10.1016/0028-3932(71)90067-4

Pelli, D. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442. doi:10.1163/156856897X00366

von Aster, M., Neubauer, A., & Horn, R. (Eds.). (2009). *Wechsler Intelligenztest für Erwachsene: Deutschsprachige Bearbeitung und Adaptation des WAIS-III von David Wechsler [Wechsler adult intelligence scale: German editing and adaptation of the WAIS-III by David Wechsler]* (2nd, corrected ed.). Frankfurt am Main, Germany: Pearson Assessment & Information.