SUPPLEMENTAL MATERIAL

Study 1

Procedure

All participants started at the first level, and advanced to higher levels when they got 2 or 3 problems correct out of a set of 3. If the participant only got 1 problem correct out of a set of 3, the next set of problems would be 1 level lower. The first problems presented (level 1) consisted of two addends that could be summed with only 1-2 changes to the abacus state. Level 2 problems also consisted of 2 addends, but required 3-4 moves on the abacus to arrive at the answer. All level 1 and level 2 problems had 2-digit answers. Level 3 problems were also 2 addends, but required 5-6 abacus moves to get the answer and could have 2 or 3 digit answers. Problems from levels 1-3 were randomly selected from lists of 60 problems each that met the listed criteria. For levels above 3, problems were randomly generated with a set number of addends. Level 4 problems had three addends, level 5 problems had four addends, and so on.

Gesture Coding

Size

The first dataset included all trials from the first and ninth minute of the task for every participant, as well as additional minutes such that each participant had at least 10 coded trials (maximum: 36 trials *M* = 21.1, *SD* = 6). In the second dataset, each trial generated its own video, allowing us to randomly select trials to code from all trials a child attempted. Between 2 and 16 trials (*M*=11.55, *SD*=2.35) were coded for each of the 143 participants. For the second dataset, coders coded a combination of baseline and spatial interference trials (conducted after this task and not included in this paper), while remaining as blind as possible to which trials were which. Only trials from the interference task were double-coded for reliability on moves measures, but we have no reason to believe that this manipulation would affect coding of minimum and maximum moves.

Number of Moves

Overall, 960 trials were coded from the first dataset by the same coders who coded size and number of moves for the second dataset. These trials were taken from the first and ninth minute of the task for every participant, as well as additional minutes such that each participant had at least 4 and up to 20 coded trials (*M* = 11.85, *SD* = 2.82). In the second dataset, the same 1665 trials that were coded for size were also coded for number of moves

Gesture Size Additional Analyses

Another variable of interest was the magnitude of the sum or of the participant's response. The problem sum is also strongly correlated with problem level ($r = .95$) but did not significantly improve the fit of the model ($p < 0.2$). The magnitude of a participant's response to the problem did significantly improve the fit of the model ($p < .05$), and significantly predicted gesture size ($\beta < 0.01$, t = 2.06, $p = .04$). Interestingly, a larger response was predictive of smaller gestures. Adding child response to the model did not change the significance of problem difficulty ($\beta = 0.05$, t = 2.33, $p = 0.02$) or log of level ($\beta = 0.19$, t = 4.12, $p < 0.01$).

Because the four coding categories are not truly a continuous variable, we repeated the above analyses treating it as a binomial variable, grouping together sizes 1 and 2 and sizes 3 and 4 into two categories. Only log of level was a significant predictor of difficulty in this model ($\beta = 1.19$, z = 4.02, $p < 0.01$).