Supplemental Data: Shortened versions (16, 20, 32, 64 items), Study/Test

Methods: The design of the shortened MST was exactly the same as Experiment 1, with the following exceptions: 1) the recognition test was a surprise (as in the incidental study condition, participants were not informed at the time of the indoor/outdoor encoding that they would later be tested on the images with an old/similar/new task); 2) we systematically varied the size of the stimulus sets among 4 conditions: 16, 20, 32, or 64 items per stimulus type (repeat, lure, foil); and 3) each participant engaged in two runs of the experiment, pseudo-randomly assigned to a pair of conditions (e.g., 16/64, 20/32, 32/32, 32/64, etc. counterbalanced for which stimulus length was presented first.). Stimulus sets were not repeated between the two runs to minimize interference from the prior run and participants were instructed that the second run was independent from the first, with no items repeated between them.

A total of 73 (mean age: 20.1; 64F/9M) young participants were divided roughly equally among seven set size conditions (16/16, 20/20, 32/32, 64/64, 16/64, 20/64, and 32/64).

Results: The data were first analyzed with a 4x2 ANOVA, using set size (16, 20, 32, or 64) and memory score (LDI and REC) (see Supplemental Table 1 for response rates in each condition). We averaged the data for each individual who performed two runs with the same set size (e.g. 16/16, 64/64) and entered into the ANOVA as a single value for each individual. We found a main effect of memory score (F(1,162) = 321.1, p<.001), with greater performance for REC (.83) than LDI (.46), but importantly, no main effect of set size (F(3,162) = 1.7, p = .17) or interaction (F(3,162) = 1.4, p = .22) (Supplemental

Figure 1). Thus, even with fewer trials per condition, the measures of the MST are consistent.

Since participants engaged in two runs of the task, we examined the effect of the first run compared to the second run. A paired t-test on the LDI scores revealed no significant difference between run order (first=0.43, second=0.47, $t(57) = 1.6$, $p>.1$). Likewise, a linear fit on the first vs. second test revealed an estimated slope of 0.98 (95% CI: 0.88 – 1.1, extra sum of squares of slope differing from 1.0: $F(1,72)=0.068$, $p=0.79$; intercept constrained at 0) suggesting very similar performance in the two runs. In contrast, a paired t-test on the REC scores revealed slightly higher REC scores for the 1st run (.84) than the 2nd run (.80; $t(57) = 2.3$, $p<.05$).

While there were no differences in LDI performance for between the different set sizes, we were interested in the reliability of this measure for each set size. As the set size is reduced, quantization error will become a factor as there are fewer and fewer trials on which to estimate the probability of key response categories (e.g., p("Similar"|Lure)). To estimate this in as unbiased a way as possible, given the design, we correlated the LDI scores for participants at a given set size (e.g., 16 items) with the other LDI scores for those participants, regardless of set size. Thus, while the variance on one side of the correlation differs across set sizes, the variance on the other side of the correlation is constant (16 vs 16-64, 20 vs. 20-64, etc.) The Pearson R^2 and corresponding p-value for each set size is presented in Supplemental Table 1. While there was a positive relationship for LDI performance for each set size, the variance in the 16-item condition prevented it from having a significant relationship with the other set sizes. We suspect that the issue for the lower correlation for the 16-item version results from quantization error, or attempting to produce a smooth, scalable variable from very few discrete trials.

Supplemental TABLE 1.

16-item	20-item	20-item	32-item
R^2 = .24 (p=.34) 95%CI (-.25, .63)	R^2 = .60 (p<.01) 95%CI (.22, .82)	R^2 = .55 (p<.01) 95%CI (.18, .78)	R^2 = .63 (p<.001) 95%CI (.41, .78)

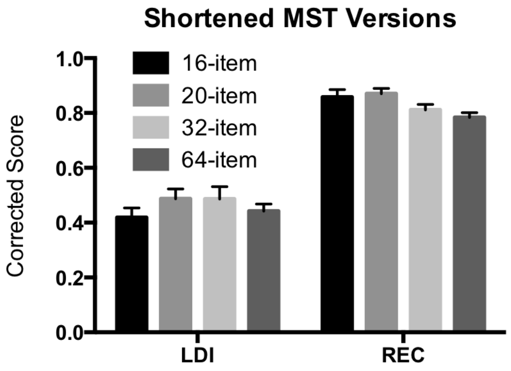
Supplemental Table 1: Correlations between each set size LDI and all other set sizes within each participant revealed reliable positive relationships for every set size, except the 16-item.

Supplemental TABLE 2.

	Repeat Old	Repeat Similar	Repeat New	Lure Old	Lure Similar	Lure New	Foil Old	Foil Similar	Foil New
16-item	.90	.09	.01	.42	.54	.04	.04	.12	.84
20-item	.88	.10	.02	.36	.56	.08	.02	.09	.89
32-item	.81	.14	.05	.34	.58	.08	.02	.10	.88
64-item	.81	.14	.05	.31	.60	.09	.03	.15	.82

Supplemental Table 2: Average percent endorsed for each trial type and each response for each of the four shortened MST versions.

Supplemental FIGURE 1.



Supplemental Figure 1: LDI and REC scores for each of the stimulus set sizes.

Supplemental TABLE 3.

	Repeat Old	Repeat Similar	Repeat New	Lure Old	Lure Similar	Lure New	Foil Old	Foil Similar	Foil New
EXP 1									
Young	.78	.16	.06	.40	.48	.12	.04	.19	.77
Aging	.79	.11	.10	.59	.28	.13	.10	.18	.72
EXP 2									
Young	.82	.11	.07	.24	.61	.15	.02	.08	.90
Aging	.82	.10	.08	.37	.42	.21	.03	.07	.90

Supplemental Table 3: Average percent endorsed for each trial type and each response for Experiments 1 and 2.

Supplemental TABLE 4.

	Repeat Old	Repeat New	Lure Old	Lure New	Foil Old	Foil New
EXP 3						
Young	.87	.13	.65	.35	.17	.83
Aging	.92	.08	.79	.21	.08	.92
EXP 4 Gist						
Young	.88	.12	.67	.33	.17	.83
Aging	.88	.12	.74	.26	.14	.86
EXP 4 Veridical						
Young	.85	.15	.33	.67	.08	.92
Aging	.87	.13	.58	.42	.07	.93

Supplemental Table 4: Average percent endorsed for each trial type and each response for Experiments 3 and 4.