Supplemental methods and results

Table S1

Raw search response times by validity and memory color

	Memory color			
Validity condition	Red	Blue	Green	Yellow
Valid	579 (133)	581 (136)	576 (129)	580 (136)
Neutral	587 (134)	587 (136)	588 (135)	589 (136)
Invalid	606 (144)	598 (144)	611 (152)	613 (151)

Note. Correct search response times in ms. Standard deviations are presented in parentheses

Table S2

Preprocessed search response times by validity and memory color

	Memory color			
Validity condition	Red	Blue	Green	Yellow
Valid	.258 (.134)	.262 (.136)	.254 (.128)	.255 (.135)
Neutral	.268 (.137)	.270 (.137)	.269 (.135)	.268 (.139)
Invalid	.292 (.152)	.283 (.148)	.297 (.157)	.296 (.159)

Note. Preprocessed response times in scaled ms. Standard deviations are presented in parentheses

S1. Memory-based attentional bias by memory color

A repeated-measures ANOVA on preprocessed response time classifier inputs (see Supplemental Table S2) across factors of validity and memory color revealed a significant main effect of validity, $F(2, 198) = 251.2$, $p < .001$, $\eta^2 = .72$, with no main effect of color. There was a significant interaction between validity and memory color, $F(6, 594) = 8.0$, $p < .001$, $\eta^2 = .07$, indicating that one possible signal that the classifiers may have learned was the differences in magnitude of mean validity effects.

To verify that the main classification signal was indeed the changing pattern of validity effects, as opposed to the magnitude of the validity effect, an additional analysis subtracted the mean validity effects from the appropriate features per memory color. (Preprocessing steps presented in the main analyses did *not* preprocess the data by validity condition, so as to maintain a decoupling between the specific validity relationships and the memory color labels.)

Within-subject classification thus tested whether these residual patterns can predict WM color. Mean decoding accuracy for distance analysis was 22.9% (*SD* = 6.5%); for logistic regression it was 27.9% (*SD* = 5.7%); and for linear SVM it was 29.2% (*SD* = 5.1%). These residual classifiers did not perform as well as the main classifiers (see Table 1), supporting our interpretation that the primary multivariate signal driving the main classification was the changing pattern of validity effects across memory colors.

S2. Memory-driven attentional bias across multiple searches

While previous dual-task studies have demonstrated group mean validity effects using a single intervening search between the WM cue and probe (e.g., Dowd, Kiyonaga, Beck, et al., 2015; Olivers, Peters, Houtkamp, & Roelfsema, 2011; Soto, Heinke, Humphreys, & Blanco, 2005), we here employed series of 12 intervening searches, in which the pattern of validity condition was randomly ordered. A repeated-measures ANOVA across validity and trial half (i.e., whether a particular search fell in the first or second half of the series) revealed main effects of validity, $F(2, 198) = 245.97$, $p < .001$, $\eta^2 = .71$, and trial half, $F(1, 99) = 8.17$, $p = .005$, $\eta^2 = .08$, such that search times were overall slower in the second half of the search series. A significant interaction effect, $F(2, 198) = 8.39$, $p < .001$, $\eta^2 = .08$, indicated that the size of the validity effect was reduced (although still significant and robust, $F(2, 198) = 121.30$, $p < .001$, $\eta^2 = .55$, through the second half of the search series (see Fig. S1). Thus, maintaining an item in WM impacted attentional bias across multiple subsequent searches, validating the inclusion of entire patterns of RTs as classifier features.

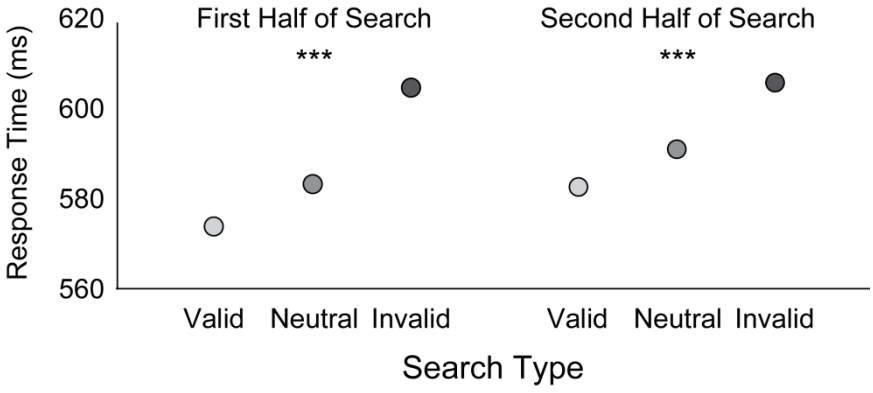

Fig. S1 Mean search response times revealed significant validity effects across the entire series of 12 searches. Validity effects were still significant after splitting the series of 12 searches into first and second

halves, indicating that a single instance of WM maintenance could impact attentional bias across multiple subsequent searches. The diameter of each *dot* represents 95% confidence intervals. *** $p < .001$

S3. Support vector machine methods

For multiclass SVM classification, we used the "e1071" package in R. This package fits a series of six binary one-against-one classifiers, resulting in decision values for all binary classifiers. A logistic distribution is fit using maximum likelihood to these decision values, and the a posteriori class probabilities for multiple classes are computed using quadratic optimization. Thus, the package outputs predicted probabilities for each of the four memory classes, and the winning probability is chosen as the winning class label. This is a standard technique for applying SVMs to multiway classification problems, and further details can be found in the documentation for the "e1071" package (Dimitriadou, Hornik, Leisch, Meyer, & Weingessel, 2006).

S4. Nonlinear feature inclusion boosts classification

Classifier inputs were normalized search response times (as organized by the 12 search combinations), supplemented with both the squares and square roots of the response time vector, resulting in a 36-features classifier input matrix. The addition of quadratic and square-root information allowed classifiers to fit potentially nonlinear relationships present in the data, just as a general linear model with quadratic and interaction terms can capture nonlinear relationships between independent and dependent variables in standard analyses. The inclusion of nonlinear features did boost classifier performance: for linear SVM, within-subject decoding accuracy with only 12 normalized response time features was 28.9% (SD = 8.2%), while within-subject decoding accuracy with all 36 features was 31.3% (SD = 5.6%). A Wilcoxon signed-rank test revealed that 12-feature SVM accuracy was significantly lower than 36-feature SVM accuracy, $Z = 3.61$, $p < .001$.

S5. Univariate approaches fail to classify WM content

Successful trial-level classification of WM contents would not be possible with the standard dual-task paradigm that combines a WM task with a single search (e.g., Soto et al.,

2005), as there is simply not enough diagnostic information within a single response time. Applying a univariate analysis to our current 12-search design also fails to classify WM content, as the data preprocessing steps explicitly removed any mean response time differences by memory color. In other words, if we were to calculate a single average response time across the 12 features for each trial, and then average these trial means by memory color within each subject, there would be no differences. Thus, our empirical results rely on a multivariate matrix of response times, which features changing patterns of validity effects.

S6. Classification subsampling analyses

It is noteworthy that our between-subject classifiers had generally higher decoding accuracy compared to within-subject approaches. A likely source of this superiority is the fact that the between-subjects analysis benefited from higher volumes of input data. We corroborated this possibility with additional subsampling analyses that examined how classifier input volume impacts classifier performance, across both within-subject and between-subject approaches (see Fig. S2). These analyses thus addressed the question of how many trials are required for obtaining above-chance decoding, and at what point adding more trials does not produce further improvement in classifier performance.

For within-subject analyses, each subject's data were randomly subsampled into approximate set sizes of 20, 40, 60, and 80 trials, across 50 repetitions. For each subsample, a linear SVM classifier was trained using a leave-one-trial-out cross-validation scheme—thus, the smallest within-subject subsample classifier was trained on 19 trials and tested on one trial, iteratively. For each subsample, we created 50 group-level accuracy vectors by randomly selecting and combining one accuracy value per subject, resulting in 50 group-level mean decoding accuracies per set size. Figure S2a shows that mean within-subject classifier performance increased with greater input volume, but leveled off at a set size of 60 trials (M = 31.1%, SD = 0.2% across 50 repetitions).

For between-subject analyses, the entire dataset was randomly subsampled into set sizes of 10, 20, 30, 50, and 100 subjects, across 50 repetitions. For each subsample, a linear SVM classifier was trained using a 10-fold cross-validation scheme—thus, the smallest between-subject subsample classifier was trained on data from nine subjects (up to 720 trials) and tested on data from one subject (up to 80 trials), iteratively. Figure S2b shows that although

between-subject classifier performance did increase with greater input volume, mean decoding accuracy was already reliably greater than chance with as few as 10 subjects (M = 33.5%, SD = 2.2% across 50 repetitions), $t(49) = 27.1$, $p < .001$, and leveled off by 50 subjects (M = 36.0%, SD = 0.2% across 50 repetitions).

Note that even the largest subsample size for the within-subject classifier (80 trials) is much smaller than the smallest subsample size for the between-subject classifier (10 subjects × 80 trials). To compare the effects of input volume on within-subject and between-subject classifier more directly, we also ran a between-subject SVM classifier with a sample size of 10 subjects, using a leave-one-subject-out cross-validation scheme, in which only eight trials were randomly pulled from each subject, balanced across memory colors. This resulted in a between-subject subsample classifier that was trained on data from 9 subjects (72 trials) and tested on data from one subject (eight trials). Across 50 randomized repetitions, mean classifier accuracy was 28.2% (SD = 4.7%). While this is not an exact comparison—likely due to variance between subjects—this result demonstrates that a between-subject classifier with decreased input volume does not outperform a within-subject classifier with similar input volume.

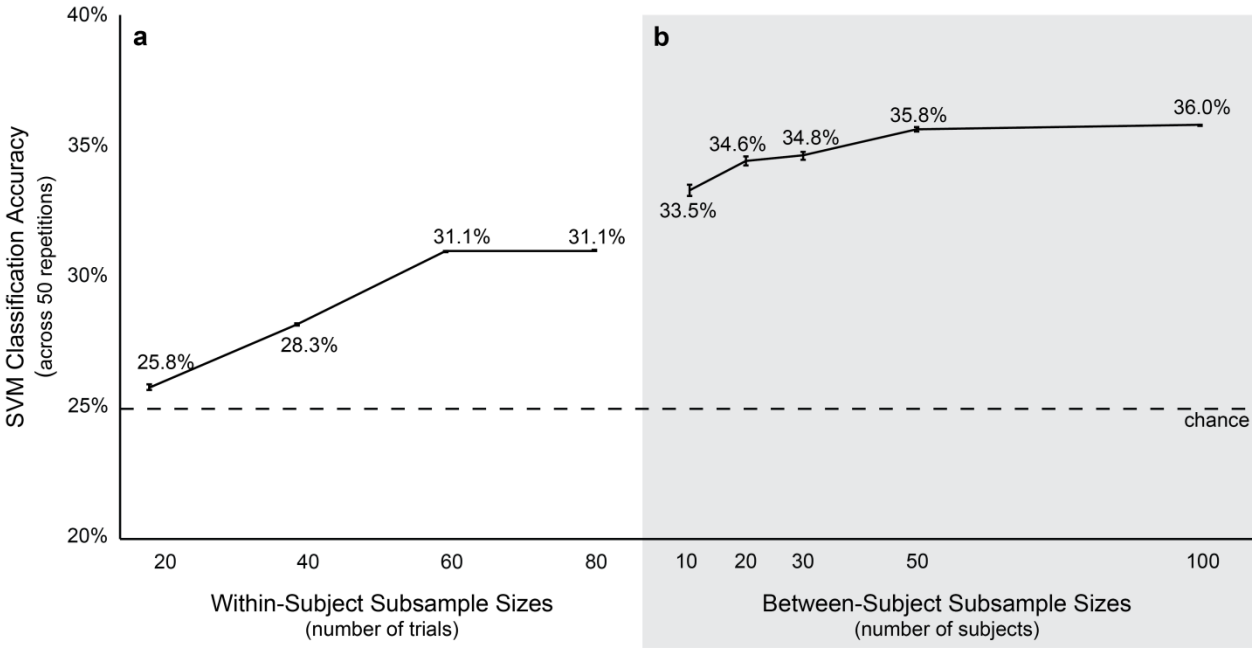

Fig. S2 Subsampling analyses revealed that linear SVM classifier performance generally increased with greater classifier input volume, across both (**a**) within-subject and (**b**) between-subject approaches. Across 50 randomized repetitions within each subsample size, within-subject (leave-one-trial-out) decoding accuracy plateaued at set sizes of approximately 60 trials, while between-subject (10-fold) decoding accuracy plateaued at set sizes of 50 subjects. Subsampling analyses also support the hypothesis that the generally higher between-subject classification performance (see Fig. 5) was an

artifact of higher volumes of input data. *Numeric means* are labeled, and *error bars* represent standard error across 50 repetitions.

S7. Within-subject classification effect sizes

Because we employ nonparametric permutation tests of significance, effect size estimates (which are impacted by departures from normality and heterogeneity of variances) are not necessarily applicable without additional assumptions. Furthermore, because between-subject classifiers output only a single decoding accuracy, it becomes more difficult to estimate effect size. Nevertheless, if we assume normality and apply standard effect size estimates, within-subject classification effect sizes would be: for distance analysis, Cohen's d = 0.71 and Hedges' g = 1.64; for logistic regression, Cohen's d = 1.14 and Hedges' g = 2.61; and for SVM, Cohen's d = 1.86 and Hedges' g = 4.23.

Supplemental references

Dimitriadou , E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2006). *e1071: Misc functions of the department of statistics (e1071).*

Dowd, E. W., Kiyonaga, A., Beck, J. M., & Egner, T. (2015). Quality and accessibility of visual working memory during cognitive control of attentional guidance: A Bayesian model comparison approach. *Visual Cognition, 23*(3), 337–356.

Olivers, C. N. L., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences*, *15*(7), 327–334.

Soto, D., Heinke, D., Humphreys, G. W., & Blanco, M. J. (2005). Early, involuntary top-down guidance of attention from working memory. *Journal of Experimental Psychology: Human Perception and Performance, 31*(2), 248–261.