Supplemental Material S4. Detailed methods and results for Bayesian model fitting, including justification of prior assumptions, and checks for convergence, sensitivity, and posterior prediction.

Gibbs Sampling

The Bayesian framework for model fitting (i.e., estimating ability values from data samples) uses probability distributions to describe our confidence in the possible values of each ability. *Prior distributions* describe our confidence in ability values before observing data, and these are transformed into *posterior distributions*, describing the likelihood of each ability value after observing data, using Bayes' theorem. To avoid complicated calculus problems, we employ a well-known approximation method known as Gibbs sampling.

All model fitting and statistical analyses were carried out in MATLAB R2014a (The MathWorks, 2014). Bayesian models were specified using the WinBUGS language (Bayesian inference Using Gibbs Sampling for Windows; Lunn et al., 2000), and Gibbs sampling was executed with the JAGS software (Just Another Gibbs Sampler; Plummer, 2003), which runs natively on Linux. The MATJAGS software (Steyvers, 2011) was used to interface MATLAB and JAGS. For Bayesian estimation of abilities, we began with assumptions about the prior distributions for baseline abilities, and changes in abilities across time were assumed to be normally distributed around zero. We then updated these assumptions with the observed data and Bayes theorem to create posterior distributions, using Markov Chain Monte Carlo integration (Gelman & Shirley, 2011).

Sampling chains were initialized with values drawn from a standard normal distribution. For the MPT model, posterior distributions for model parameters were constructed using four chains of 2,000 samples after 1,000 burn-in samples, yielding a total of 8,000 samples of the posterior distribution of each participant ability. For the IRT model, posterior distributions for model parameters were constructed using four chains of 1,000 samples after 500 burn-in samples, yielding a total of 4,000 samples of the posterior distribution of participant ability.

We constructed posterior samples of the IRT-P(S) change statistic by first converting each sample of the latent accuracy parameters at each testing session into a probability of success on an item of average difficulty, then taking the difference between the corresponding samples of the two testing sessions, yielding a new chain of posterior samples for the change in IRT-P(S). We constructed posterior samples of the MPT-P(S) change statistic by first converting each sample of the ability parameters at each testing session into a probability of success on an item of average difficulty, then taking the product of all probabilities at corresponding samples (i.e., within each single sample of posterior ability parameters, without contamination across different samples in the chain) to produce a sample of MPT-P(S) on a given testing session, and finally taking the difference between the corresponding samples of the two testing sessions, yielding a new chain of posterior samples for the change in MPT-P(S). A similar procedure was undertaken to produce posterior estimates of changes in MPT-E(D), first calculating MPT-E(D) for each sample of ability parameters at a given testing session, and then taking the difference between corresponding samples of two different testing sessions to produce a sampling chain of the changes in MPT-E(D).

The means of the posterior samples were taken as point estimates of the summary statistics. Credible intervals were constructed by sorting the posterior samples for the quantities of interest and taking the 2.5 and 97.5 percentile samples as lower and upper bounds.

Convergence Check

For the MPT model, across both treatment groups, we estimated a total of 864 participant variables directly (i.e., the participant abilities at each testing session), and we were interested in the derived variables: MPT-P(S), MPT-E(D), and changes in these variables between testing sessions at baseline and immediately following completion of therapy (i.e., Weeks 1 and 12). For each of these variables, longer posterior sample chains (4,000 samples each) were initially plotted and visually inspected for convergence, to estimate a reasonable number of burn-in and retained samples (Lee & Wagenmakers, 2014). After determining the default sampling parameters, we examined the potential scale reduction factor ($\hat{R}$) for each estimated variable; values below 1.05 were considered satisfactory (Gelman & Shirley, 2011). The same convergence checks were applied to the IRT model estimates of ability.

Results. Convergence was satisfactory for all IRT model-based quantities of interest (max. $\hat{R}$ = 1.006). Convergence was also satisfactory for all MPT model-based quantities of interest (max. $\hat{R}$ = 1.003).

Justification of Default Prior Assumptions

The default prior assumptions were made with the intention of minimizing bias in the posterior estimates of the summary statistics. This goal was motivated by the *objective Bayes* approach to defining prior distributions, wherein as few assumptions are made as possible (Rouder et al., 2009). For the IRT model, due to the known item heterogeneity in the model, prior expectations about the latent probability of a correct response are shifted upward or downward based on the known difficulty of the item, meaning that we can only be minimally informative about the probability of a correct response on the average difficulty item (or some other arbitrary point on the item difficulty scale). To do this, we set the mean of the logit-normal prior distribution (i.e., a normal distribution on the logit scale ranging [-inf, +inf]) equal to the average item difficulty (-0.18), and we set the standard deviation (i.e., the dispersion) to be 1.7, to approximate a uniform distribution on the probability scale.

Due to the structure of the MPT model, there is a complex relationship between the parameters of the prior logit-normal distributions for each of the six latent abilities and the prior distribution for the probability of a correct response. We used simulations to explore the effects of different logit-normal parameters, taking six random draws, converting them to probabilities based on the average difficulty values of the test items, and then taking their product to yield a single sample of the prior distribution for the probability of a correct response; this sampling procedure was repeated 10^7 times to produce simulated prior distributions (Supplementary Figure S5, top row). It was determined that a mean of 2.5 and standard deviation of 2.8 for all latent abilities produced a reasonably diffuse distribution for the probability of a correct response. While these values are somewhat arbitrary (i.e., there are other prior assumptions that could produce an approximately similar distribution on the probability scale), with enough observed data and reasonably plausible prior assumptions, Bayesian estimation and inference is largely insensitive to changes in the prior assumptions, as we demonstrate in the sensitivity analysis described in the next section.

Finally, in both models, the longitudinal changes in abilities were assumed to vary normally around zero, motivated by the default hypothesis that no real change in the latent abilities occurred; to the extent that posterior distributions are shifted away from this prior location after observing data, the evidence will support the claim that a real change in abilities did occur. Note also that the dependence of later abilities on earlier abilities implements a

longitudinal regime that constrains model estimates based on the time-structure of the data, rather than assuming that each observation in time is independent from the others.

Sensitivity Check

We did not investigate the sensitivity of the IRT model parameter estimates to the prior specifications. However, for the MPT model, in addition to the default prior assumptions that were minimally informative about the probability of a correct response at baseline, we examined two other prior specifications for baseline abilities, as well as a different random seed for the default specification. First, we specified prior distributions that were minimally informative about the probability of success on a latent process with average difficulty. Like the Bayesian IRT model, due to the known item heterogeneity, prior expectations about the probability of success on latent processes are shifted upward or downward based on the known difficulty of the process, meaning that we can only be minimally informative about a single point on the process difficulty scale; we chose the average difficulty of all processes as our anchor point. Thus, we set the mean of the prior logit-normal distribution to the average difficulty over latent processes (-1.8), and we set the standard deviation to 1.7 to approximate a uniform distribution on the probability scale (Supplementary Figure S5, second row). Second, we specified informed prior distributions based on the descriptive statistics of point estimates from the 90 independent participants reported in Walker et al. (2018) who were examined by the same research group that conducted the current study (Supplementary Figure S5, bottom row). Sensitivity checks were conducted using only the data from Treatment Group 1.

Results. The coefficients of determination (R^2) between parameter point estimates and interval estimate widths obtained using different random seeds or specifications of the prior distributions were all greater than .98. These results indicate that the point estimates and interval widths that are used for inference can be obtained using a variety of reasonable starting assumptions.

Posterior Predictive Check

To assess the MPT model's fit to the data, we performed posterior predictive checks and examined posterior predictive error. The logic of the posterior predictive check is that a model's estimated parameters specify some distribution of possible data that might be observed, and, ideally, the data that are actually observed should appear to be a reasonable sample from that distribution. To calculate a posterior predictive *p*-value, a random draw is taken from the posterior distribution of parameter values (i.e., a posterior sample from the sampling chain used for fitting); new data are randomly generated from the model using these parameter values; and the prediction error for this new data (i.e., the difference between the data's response type frequencies and the model's expected values) is compared with the prediction error for the observed data. This procedure is repeated for each posterior sample in the chain, and the proportion of new data samples that are closer to the model's predicted values than the original data defines the posterior predictive *p*-value. A low posterior predictive *p*-value means that the original data do not fit the model's predictions as well as data sampled directly from the model; however, this does not necessarily mean that the model's fitted parameters are useless for making inferences about participants, which requires additional investigations to determine. Posterior predictive *p*-values less than .05 were taken as indicators of poor recovery of the response frequencies from the model parameters. We generated posterior predictive *p*-values based on the T_1 test statistic proposed by Klauer (2010) for examining recovery of mean

response type frequencies. We tested the recovery of participant-wise response frequencies summed over items, for each response type separately and for the response type distribution as a whole, for each naming test.

We also examined the posterior predictive error of the fitted model: For individual response types, we examined the absolute difference between the observed response type proportions in the data and the predicted response type proportions from averaging over samples of the posterior predictive data that is generated from the model. For response type distributions, we examined the root mean square error between all observed and predicted response type proportions.

Results. Regarding the fit to response type distributions generated by participants (considering all response types together), there were three tests (4%) in Treatment Group 1 with posterior predictive $p < .05$, and there were zero tests (0%) in Treatment Group 2 with posterior predictive $p < .05$. This means that nearly all the model's prediction accuracies for the observed distributions of response-type rates were similar to the model's prediction accuracies for new data that was generated directly from the model. The median RMSE between predicted and observed distributions of response rates was 0.6%; the maximum RMSE was 3.3%. Recall that the RMSE is the expected prediction error for a randomly selected response type. For comparison, when fitting the Foygel and Dell (2000) spreading activation model to response-type distributions, Schwartz et al. (2006) reported an average RMSE of 2.4%, with 23 of their 94 participants having RMSE greater than the maximum RMSE reported here for the MPT model.

With regard to specific response types, there were no tests in either treatment group with posterior predictive $p < .05$ for Correct, Semantic, Formal, Neologism, or No Attempt response types, indicating that the prediction accuracies for these observed response rates were similar to prediction accuracies for response rates generated directly from the model. For Mixed errors, there was a single test (1%) in Treatment Group 1 that had posterior predictive $p <.05$. For Unrelated errors, there were 23 tests (30%) in Treatment Group 1 with posterior $p < .05$, and there were 24 tests (35%) in Treatment Group 2 with posterior $p < .05$. For Abstruse Neologism errors, there were 19 tests (25%) in Treatment Group 1 with posterior $p < .05$, and there were 12 tests (18%) in Treatment Group 2 with posterior $p < .05$. Even though prediction errors were sometimes larger than expected if the model had been true for these response types, the median prediction error for any response rate was less than 1%; the maximum prediction error for any response rate in any participant was 6.1%, for Abstruse Neologisms. The predicted rates of Abstruse Neologism and Unrelated errors strongly depend on assumptions about the probability of a phonological error resulting in a real word. Because these probabilities were estimated from an independent cohort with much lower rates of speech motor impairments, these probabilities may have been overestimated. Nevertheless, within the bounds of the model's current assumptions, the inadequate model fits in this minority of cases do not by themselves suggest that the parameter estimates are biased or useless. Because the overall model fit was deemed generally adequate for most participant-level response types and distributions, and to investigate the general utility of the estimated parameters and summary measures, no participants or items were excluded in the subsequent analyses based on posterior predictive checks.