Assuming independence between the choice alternatives and independence between the features contained within each alternative, the ROUSE model established features as the basic unit of evidence evaluation (Huber et al., 2001). In the original formulation, feature likelihood ratios (i.e., the likelihood that the feature belonged to the target) were determined for situations in which features were either active, or inactive (i.e., binary valued) and features were either known to exist in a prime or not (i.e., certain knowledge for the primes). These four likelihood ratios were calculated assuming three potential sources of independent feature activation ($\alpha$ = activation from presentation of the primes, $\beta$ = activation from the brief presentation of the target, and $\gamma$ = noise activation).

The present reformulation of ROUSE in terms of a Bayesian belief net reworks the feature likelihood ratio calculations, unifying the four expressions into a single equation. This is done in order to understand the behavior of ROUSE when feature activation and prime knowledge are probabilities, rather than dichotomous variables. If these probabilities are known values, rather than parameters driving a stochastic process with dichotomous results, a mapping is possible between the original ROUSE model, and the graded activation dynamics with synaptic depression.

The belief net seen in Fig. 5 demonstrates the situation. In the real world (the activation portion of the figure), a feature is activated by either the prime, with probability $\alpha$, or the target, with probability $\beta$. If neither source exists, the feature is nevertheless active with probability $\gamma$. This noise activation could be included as a separate causal link in Fig. 5, although, assuming that noise is always present and independent of the other sources, Equations A7-A10 appropriately factor in noise activation. Whether the prime and target sources actually exist depends on whether the

feature has been primed and whether the feature is contained in the target. This reformulation presents ROUSE as a generative model, using the same causal structure to activate features and to infer the probabilities that the prime and target exist as sources based upon the observed feature activation. It is assumed that the real activation state (the feature below the dashed line) is copied into the inference activation state (the feature above the dashed line). Crucially, it is assumed that the inference process does not have access to the true activation probabilities and must use potentially inaccurate estimates of these probabilities (i.e., $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$).

In keeping with the original ROUSE model, we first calculate likelihood ratios, although any likelihood ratio is easily transformed into a corresponding probability. The likelihood ratio, $\lambda(T)$, is the posterior probability that the target is a source, $p(T)$, divided by the posterior probability that the target is not a source, $p(\overline{T})$,. We initially calculate two separate likelihood ratios, conditioning on the feature existing in an active state, $F$, Equation A1, or an inactive state, $\overline{F}$, Equation A2.

$$\lambda\left(T \mid F\right) = \frac{p(T \mid F)}{p(\overline{T} \mid F)} \tag{A1}$$

$$\lambda\left(T \mid \overline{F}\right) = \frac{p(T \mid \overline{F})}{p(\overline{T} \mid \overline{F})} \tag{A2}$$

Next, the conditional probabilities are reversed, using Bayes theorem, and the common denominator terms are removed, resulting in Equations A3 and A4 in which $T_o$ is the prior probability of the target existing as a source.

$$\lambda\left(T \mid F\right) = \frac{T_o p(F \mid T)}{\overline{T}_o p(F \mid \overline{T})} \tag{A3}$$

$$\lambda\left(T \mid \overline{F}\right) = \frac{T_o p(\overline{F} \mid T)}{\overline{T}_o p(\overline{F} \mid \overline{T})} \tag{A4}$$

These equations are expanded by additionally conditioning on whether the prime is, $P$, or is not, $\overline{P}$, a potential source of activation, with each possibility appropriately weighted by the probability of the prime existing as a source, given the state of the target.

$$\lambda\left(T \mid F\right) = \frac{T_o}{\overline{T}_o}\left[\frac{p(P \mid T)p(F \mid TP) + p(\overline{P} \mid T)p(F \mid T\overline{P})}{p(P \mid \overline{T})p(F \mid \overline{T}P) + p(\overline{P} \mid \overline{T})p(F \mid \overline{T}\overline{P})}\right] \tag{A5}$$

$$\lambda\left(T \mid \overline{F}\right) = \frac{T_o}{\overline{T}_o}\left[\frac{p(P \mid T)p(\overline{F} \mid TP) + p(\overline{P} \mid T)p(\overline{F} \mid T\overline{P})}{p(P \mid \overline{T})p(\overline{F} \mid \overline{T}P) + p(\overline{P} \mid \overline{T})p(\overline{F} \mid \overline{T}\overline{P})}\right] \tag{A6}$$

However, the graph tells us that the prime and target are independent when the state of the feature is unknown (e.g., p(P | T) = p(P)), and, therefore, the probability of prime terms need not condition on the state of the target (note that this does not hold true if the decision factors in knowledge regarding how often the target is primed, although such a situation implies a different graph, including a link between the prime and target).

Assuming that the prime and target activate the feature in an independent manner, the four conditional probabilities necessary to solve Equations A5 and A6 are calculated from the estimated activation probabilities. These are most easily expressed as the probabilities that the feature is inactive (i.e., the sources of activation have all failed), resulting in Equations A7-A10. The corresponding equations for the probability that the feature is active are simply 1.0 minus each of these equations.

$$p(\overline{F} \mid \overline{T}\overline{P}) = 1 - \hat{\gamma} \tag{A7}$$

$$p(\overline{F} \mid T\overline{P}) = (1 - \hat{\beta})(1 - \hat{\gamma}) \tag{A8}$$

$$p(\overline{F} \mid \overline{T}P) = (1 - \hat{\alpha})(1 - \hat{\gamma}) \tag{A9}$$

$$p(\overline{F} \mid TP) = (1-\hat{\alpha})(1-\hat{\beta})(1-\hat{\gamma}) \tag{A10}$$

After substituting Equations A7-A10 (and 1.0 minus these equations), and performing algebraic reduction and simplification, Equations A5 and A6 become Equations A11 and A12.

$$\lambda\left(T \mid F\right) = \frac{T_o}{\overline{T}_o}\left[\frac{\hat{\beta} + p(P)\hat{\alpha} - p(P)\hat{\alpha}\hat{\gamma} + \hat{\gamma} - p(P)\hat{\alpha}\hat{\beta} + p(P)\hat{\alpha}\hat{\gamma}\hat{\beta} - \hat{\gamma}\hat{\beta}}{p(P)\hat{\alpha} - p(P)\hat{\alpha}\hat{\gamma} + \hat{\gamma}}\right] \tag{A11}$$

$$\lambda\left(T \mid \overline{F}\right) = \frac{T_o}{\overline{T}_o}(1-\hat{\beta}) \tag{A12}$$

If the prior probability that the target is a source, $T_o$, equals one half, such as is with 2-AFC testing, Equation A11 and A12 reduce to the original four feature likelihood ROUSE equations, provided that the probability of the prime existing as a source, p(P), is set to the extreme values of 0.0 or 1.0. Huber et al. (2001) demonstrated that the estimate of prime activation, $\hat{\alpha}$, critically determines whether priming results in a preference for or against primed alternatives. Underestimates of $\alpha$ result in too little discounting of primed features whereas overestimates of $\alpha$ result in too much discounting of primed features. As seen in Equation A11, every term that contains the estimate of $\alpha$ is also multiplied by the probability that the prime is a source, p(P). Therefore, this suggests an alternative interpretation of discounting. Rather than over or underestimating $\alpha$, it may be that the system always overestimates $\alpha$, but in some circumstances, such as with short prime durations, it is uncertain whether the feature has been primed (i.e., p(P) is low). Distinguishing between these interpretations is extremely difficult. Nevertheless, this reformulation makes it clear that the discounting referred to by Huber et al. is equivalent to the phenomenon referred to as "explaining away" in the belief net literature. If the

feature is active, the basic question is whether this activation is due to the prime or the target. As the probability that the prime is a source, p(P), increases, the feature activation is explained away, and the probability that the target is a source decreases.

Next, we continue with our proof in an attempt to include a term for the probability of the feature being active. We can convert Equations A11 and A12 into conditional probabilities using Equations A13 and A14.

$$p(T \mid F) = \frac{\lambda(T \mid F)}{1 + \lambda(T \mid F)} \tag{A13}$$

$$p(T \mid \overline{F}) = \frac{\lambda(T \mid \overline{F})}{1 + \lambda(T \mid \overline{F})} \tag{A14}$$

At this point, we would like to combine these conditional probabilities using a term for the probability of the feature being active, p(F). This can be accomplished with Equation A15. However, in the first step of deriving the likelihood ratios we included a term for the priors of the target. These priors on the target are theoretically the same probability as the one calculated in A15. Although Equation A15 is correct on its own, according to probability theory it is incorrect to plug in our results from Equations A13 and A14, which include the priors of the target. Nevertheless, we are left with no alternative. Indeed, for this causal graph, any attempt to factor the joint probability distribution into the fundamental generative conditional probabilities (i.e., Equations A7-A10), necessitates a term for the probability of the target and another for the probability of the prime. This means that there can be no exact posterior probability of the target except when conditioning on the state of the feature. Instead, we define the calculation performed in A15 as the expected posterior probability of the target, keeping separate the outcome of this calculation and the target priors used in calculating Equations A13 and

A14. Essentially, this calculation gives us a method for smoothly mapping between the extremes of Equation A13 and Equation A14.

There are two situations in which including priors for the target in Equation A15 can be viewed as accurate. The first assumes a population of identical belief nets, which have their feature values fixed in the present or absent state, as stochastically determined by p(F). The calculation in A15 is the expected value for p(T), conditional on whatever feature value is observed for the randomly sampled belief net. The second situation turns A15 into an exact conditional probability. For this situation, we add a new node called O, for observation, which has a directed link from F. If the feature is present, it will result in an observation with some probability, and, if the feature is absent, it will result in an observation with some lesser probability. Similar to signal detection theory, a probability that the feature is present can be inferred, based upon the observation. For such a situation, the probability of the target is now conditional on the observation, and the probability of the feature is also conditional on this observation. Because the calculation is conditional on the observation, it is allowable to include a term for the priors of the target in equation A15.

$$E[p(T)] = p(F)p(T \mid F) + p(\overline{F})p(T \mid \overline{F}) \tag{A15}$$

Finally, combining across Equations A11 thru A15, results in Equation A16, which is a general expression for the expected posterior probability that the target exists as a potential source of the feature.

$$E[p(T)] = \frac{T_o}{1 - T_o\hat{\beta}}\left[\frac{-p(F)\overline{T}_o\hat{\beta} + p(P)\hat{\alpha}\left\{T_o\hat{\beta} - T_o\hat{\beta}\hat{\gamma} + \hat{\gamma} - 1 + T_o\hat{\beta}^2\hat{\gamma} - T_o\hat{\beta}^2 - \hat{\beta}\hat{\gamma} + \hat{\beta}\right\} + T_o\hat{\beta}\left\{\hat{\gamma} - 1 - \hat{\beta}\hat{\gamma} + \hat{\beta}\right\} + \hat{\beta}\hat{\gamma} - \hat{\gamma}}{p(P)\hat{\alpha}\left\{T_o\hat{\beta} - T_o\hat{\beta}\hat{\gamma} + \hat{\gamma} - 1\right\} + T_o\hat{\beta}\left\{\hat{\gamma} - 1\right\} - \hat{\gamma}}\right] \tag{A16}$$

Equation A16 includes real valued levels of discounting, p(P), and real valued levels of feature activation, p(F). This allows us to relate the probability of feature activation to neural activation and the probability that the prime is a source to the level of synaptic depletion. In the text this is done in Figure 6, in which post-synaptic output is related to the expected posterior probability of the target existing as a source, E[p(T)]. Essentially, through synaptic depression, the synapse is calculating the probability that activation is due to a new input (the target), rather than a previous input (the prime).