Journal of Mathematical Psychology 90 (2019) 47–60

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

How many dimensions underlie judgments of learning and recall redux: Consideration of recall latency reveals a previously hidden nonmonotonicity☆

Yoonhee Jang [a,*], Heungchul Lee [b], David E. Huber [c]

[a] Department of Psychology, University of Montana, United States
[b] Net Intelligence & Research, Republic of Korea
[c] Department of Psychological and Brain Sciences, University of Massachusetts, Amherst, United States

HIGHLIGHTS

- Monotonic state-trace solutions can arise from particular measures and manipulations.
- More than one latent variable is primarily needed to explain JOLs and recall latency.
- A single latent variable is generally needed to explain JOLs and recall accuracy.
- Delayed JOLs and test practice support a multidimensional account of JOLs and recall.
- We reaffirm no evidence of a multidimensional model using intrinsic/extrinsic cues.

ARTICLE INFO

Article history:
Received 1 May 2018
Received in revised form 17 September 2018
Available online 24 December 2018

Keywords:
State-trace analysis
Judgments of learning
Recall accuracy and latency

ABSTRACT

Jang and Nelson (2005) used state-trace analysis to examine factors that affect judgments of learning (JOLs) given as a prediction of future cued recall success. Koriat's (1997) cue-utilization framework predicted that intrinsic cues (e.g., item difficulty) would have approximately the same effects on recall as they would have on JOLs whereas extrinsic cues (e.g., number of presentations) would have greater effects on recall than on JOLs. In contradiction to the prediction from the cue-utilization framework, Jang and Nelson repeatedly found a monotonic state-trace solution, suggesting that a single latent variable (e.g., memory strength) explained both JOLs and recall. However, performance can be measured in many ways, and dissociations between JOLs and recall may arise from factors other than intrinsic and extrinsic cues. Thus, an apparent monotonic solution could be an artifact of the particular choice of behavioral measures or experimental manipulations. In light of this possibility, we reanalyzed Jang and Nelson's data and conducted a new experiment, considering recall latency as well as recall accuracy and including the manipulation of immediate versus delayed JOLs. Even when additionally including both immediate- and delayed-JOL conditions, state-trace analysis with JOL magnitude and recall accuracy generally suggested a single latent variable, except for a single case in which immediate JOLs produced high overconfidence. However, the state-trace results with JOL magnitude and recall latency primarily revealed a nonmonotonic function, indicating that more than one latent variable is needed to explain the relationship between JOLs and recall.

This research concerns the processes underlying metacognitive confidence judgments and memory retrieval. Particularly, we take a step toward understanding the theoretical structures of judgments of learning (JOLs) in recall memory paradigms. JOLs are generated after study as predictions of future memory performance for studied items. For instance, a participant may study a pair of words, followed by presentation of one word as a cue to which a JOL is given, indicating the likelihood that the associated target word will be recalled on a future cued recall test. From an educational standpoint, JOLs are analogous to a student's assessment of whether they have studied sufficiently for an upcoming exam. JOL accuracy (aka, 'resolution') for the relative recallability of different

☆ We thank two anonymous reviewers and John Dunn for valuable comments on an earlier version of this article; John Dunn for his help in programming; and Mike Kalish for providing the CMR algorithm code.

* Correspondence to: Department of Psychology, University of Montana, 32 Campus Drive, Missoula, MT 59812-1584, United States.
E-mail address: yoonhee.jang@umontana.edu (Y. Jang).

https://doi.org/10.1016/j.jmp.2018.10.006
0022-2496/

studied items (i.e., accurately predicting which targets will or will not be recalled), is usually measured with the Goodman–Kruskal gamma correlation between JOLs and recall (see Nelson, 1984, for details; although see Jang, Wallsten, & Huber, 2012). Typically, the gamma correlation is extremely high when JOLs are given after a delay (at least 30 s) whereas it is low to only moderate when JOLs are given immediately after study. This difference in JOL accuracy is called the delayed-JOL effect (Nelson & Dunlosky, 1991), and it is one of the most robust findings in the literature (see Rhodes & Tauber, 2011, for a meta-analysis). A gamma correlation uses JOL magnitude and recall accuracy jointly on an item-by-item basis to quantify the correspondence between JOLs and recall performance in each condition separately. By contrast, here we relate JOL magnitude to recall performance (aka, 'calibration') using state-trace analysis across conditions (a condition-by-condition basis) in that if JOLs are sensitive to the manipulation that affects recall, then increases in recall accuracy should correspond to increases in JOL magnitude.

Specifically, the present study reports an application of state-trace analysis (Bamber, 1979; also see Dunn & Kirsner, 1988; Loftus, 1978) as a tool for illustrating the underlying processes of JOLs and recall. State-trace analysis compares different theoretical structures that may underlie changes in two (or more) dependent variables, contrasting a single-dimensional model in which dependent measures reflect the same latent variable, versus a multidimensional model in which multiple latent variables can load onto the dependent variables to different degrees. The single-dimensional model necessarily predicts a monotonic function for a state-trace plot of the two dependent variables whereas the multidimensional model can accommodate nonmonotonic functions. In the present study, we often consider specific two-dimensional models, but more generally, state-trace analysis asks whether more than one dimension is needed to capture the data. This analysis technique has proven useful in a variety of research areas, including the confidence–accuracy relation in recognition memory (e.g., Busey, Tunnicliff, Loftus, & Loftus, 2000; Heathcote, Freeman, Etherington, Tonkin, & Bora, 2009), the face-inversion effect (e.g., Loftus, Oberg, & Dillon, 2004), models of category learning (e.g., Dunn, Newell, & Kalish, 2012; Newell, Dunn, & Kalish, 2010), remember–know judgments (e.g., Dunn, 2008), and metacognitive confidence judgments, or JOLs (e.g., Jang & Nelson, 2005).

Although the demonstration of dissociations is often taken as evidence for a multidimensional model, this conclusion is not necessarily valid and might arise from the properties of the measurement scales used to evidence a dissociation (see Dunn & Kirsner, 1988; Loftus, 1978, for details). By contrast, a nonmonotonic result in a state-trace analysis provides compelling evidence of multidimensionality. In light of these concerns, Jang and Nelson (2005) used state-trace analysis to assess whether multiple latent variables are necessary to affect the correspondence between JOL magnitude and recall accuracy. If JOLs are made primarily on the basis of target retrieval, then high JOL accuracy (near-perfect gamma correlations) should occur, yielding a monotonic function between JOLs and recall, in support of the single-dimensional model. By contrast, if JOLs are affected by information that is irrelevant to eventual memory performance, then manipulations of this information should produce a nonmonotonic function, in support of a multidimensional model.

The remainder of this article is organized as follows. First, we summarize the logic and findings of a previous application of state-trace analysis conducted by Jang and Nelson (2005) in which a particular two-dimensional model was compared to the single-dimensional model for the study of JOLs. Second, we address the possibility of a multidimensional account of JOLs and recall (not just limited to the two-dimensional model) considering different manipulation variables and dependent measures, and we motivate reanalysis of Jang and Nelson's data. Third, we briefly describe a new experiment, which allows us to analyze state-trace plots in a different situation (apart from a delay after study), which improves JOL accuracy. Fourth, we report the state-trace analysis results of the new data as well as the reanalysis results of Jang and Nelson's, testing the statistical reliability of the conclusion as regards dimensionality. Finally, we discuss the resulting implied conclusions about JOLs and recall.

Jang and Nelson's test of the cue-utilization framework

Producing a qualitative conclusion from a nonparametric analysis of the data, state-trace analysis is ideally suited for testing theories even if those theories are not instantiated with a mathematical/computational model. For example, Jang and Nelson (2005) conducted a series of state-trace analyses to examine the prediction from Koriat's (1997) cue-utilization framework regarding JOLs and recall. The cue-utilization framework assumes that people assess various cues that are differentially predictive of subsequent recall, and a major distinction is made between intrinsic and extrinsic cues. Intrinsic cues, such as item difficulty and item relatedness, exert similar effects on JOLs and recall whereas extrinsic cues, such as the number of presentations and study duration, affect JOLs less strongly than recall. The distinction between intrinsic and extrinsic cues leads to the prediction of a particular two-dimensional model (in the Discussion, we consider other two-dimensional models). Fig. 1 illustrates predictions of a default single-dimensional model, which assumes that a single variable, such as 'memory strength', underlies both JOLs and recall (Panels A to C) versus the two-dimensional model predicted by the cue-utilization framework (Panels D to F). For each model, item difficulty (difficult versus easy word pairs) and the number of presentations (one versus two) are manipulated as the independent variables of intrinsic and extrinsic cues, respectively, and JOLs and recall are the two dependent variables. The first two panels of each row illustrate the separate effects of the independent variables on the two dependent variables of JOLs and recall. The prediction of each model is achieved by combining the two plots into a state-trace plot as shown in the third panel of each row. As illustrated in Panel C, the critical prediction of the single-dimensional model is that the one-presentation and two-presentation curves lie along a single monotonically increasing curve: the data shown between the two arrows in Panel C are the overlapping portion of the one-presentation curve and the two-presentation curve. By contrast, the two-dimensional model from the cue-utilization framework predicts that the two curves are separated (i.e., a curve going through all of the data would be nonmonotonic), as shown in Panel F. In this case, the two-presentation curve falls to the right of the one-presentation curve because item difficulty (intrinsic cues) has approximately the same effect on recall as it has on JOLs (one dimension for both) whereas the number of presentation (extrinsic cues) has a greater effect on recall than on JOLs (a second dimension needed for recall).

Using state-trace plots, Jang and Nelson (2005) investigated the relative contributions of intrinsic and extrinsic cues for immediate and delayed JOLs. In five experiments, Jang and Nelson used paired associate learning with JOLs given in response to the cue word alone and cued recall testing. Each experiment included one intrinsic cue (either difficult versus easy word pairs, or unrelated versus related word pairs) and one extrinsic cue (either one versus two presentations, or short versus long study duration) as the two independent variables, as summarized in Table 1 (Experiments 1A to 1D, and 2). One half of the pairs in each condition received immediate JOLs, and the other half received delayed JOLs. In an apparent contradiction to the prediction from the cue-utilization framework, all 10 state-trace plots (i.e., five experiments × two

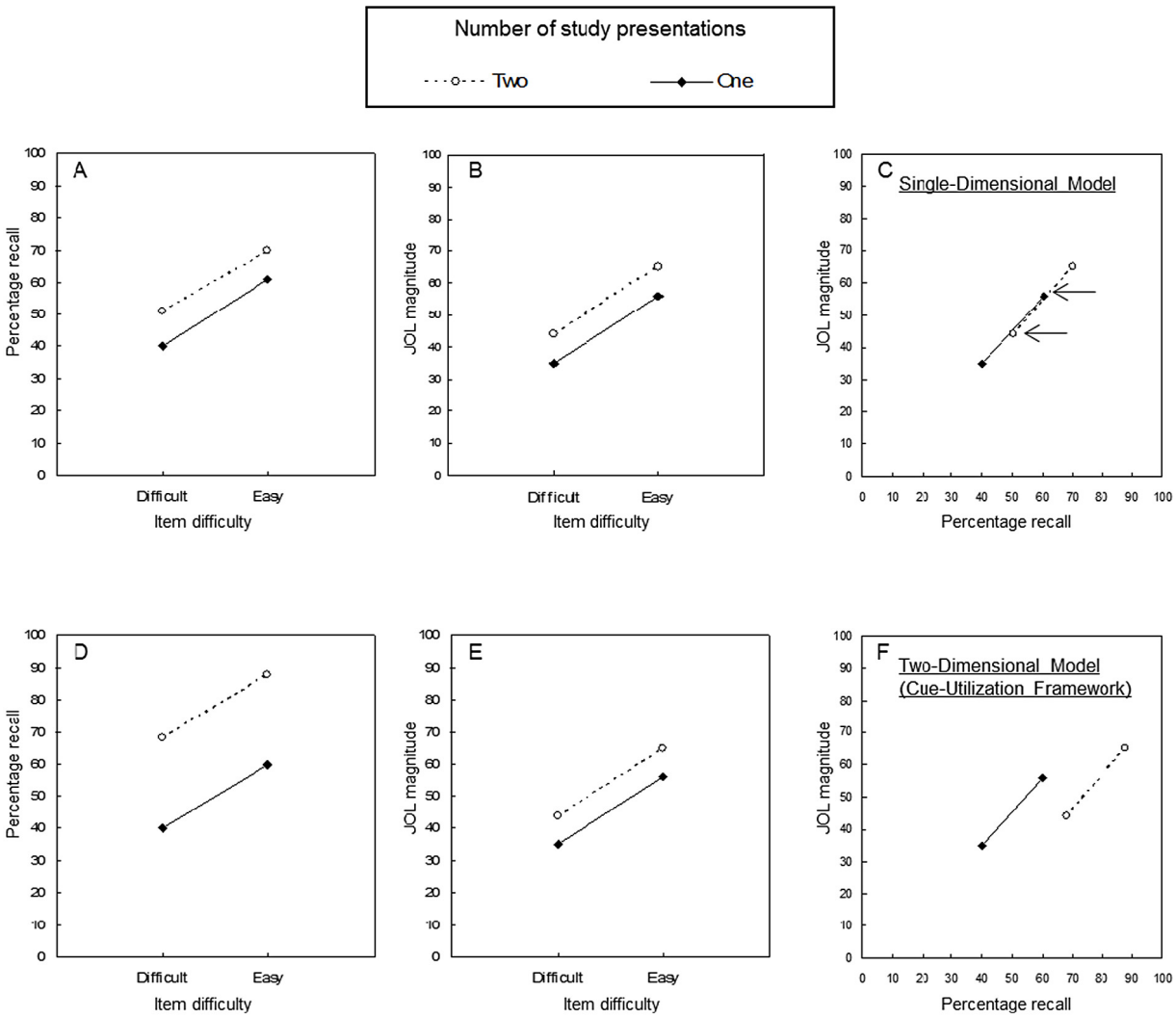

Fig. 1. Predictions of the single-dimensional model (Panels A to C); and the two-dimensional model from Koriat's (1997) cue-utilization framework (Panels D to F). Panels A, B, D, and E show traditional data in which the two dependent variables (recall accuracy and JOL [judgment of learning] magnitude) are plotted as functions of the two independent variables (item difficulty and number of presentations). Panels C and F show state traces in which JOL magnitude is plotted against recall accuracy.
Source: Adapted from Fig. 3 in "How many dimensions underlie judgments of learning and recall? Evidence from state-trace methodology", by Y. Jang, and T. Nelson, 2005, *Journal of Experimental Psychology: General, 134*, p. 311.

Table 1
Intrinsic and Extrinsic Cues Used in Experiments 1A to 1D and 2 of Jang and Nelson (2005), and the New Experiment on the Testing-JOL Effect.

Experiment	Intrinsic cue	Extrinsic cue
1A	Item difficulty: Difficult vs. easy (Swahili–English word pairs)	Number of presentations: One vs. two
1B	Item relatedness: Unrelated vs. related (concrete noun pairs)	Number of presentations: One vs. two
1C	Item difficulty: Difficult vs. easy (Swahili–English word pairs)	Study duration: Short (5 s) vs. long (15 s)
1D	Item relatedness: Unrelated vs. Related (concrete noun pairs)	Study duration: Short (2 s) vs. long (8 s)
2	Item relatedness (with instructions to compare degree of relatedness): Unrelated vs. related (concrete noun pairs)	Number of presentations: One vs. two
New	Item relatedness: Unrelated vs. related (concrete noun pairs)	Number of presentations: One vs. two through two cycles of SJT (i.e., no testing on the first SJT vs. testing on the second SJT)

Note. S = Study; J = Judgments of learning; T = Test.

JOL conditions) revealed a monotonic function, with quite small bidirectional standard errors. These results suggested that the same latent variable (e.g., memory strength) underlying recall performance was used when making JOLs. These results were somewhat surprising, not only because Jang and Nelson failed to find evidence for a multidimensional account of JOLs and recall (including the two-dimensional model from the cue-utilization framework), but also because the single-dimensional account was found regardless of immediate and delayed JOLs. Yet, most theories of the delayed-JOL effect appeal to factors other than a unitary trace strength: i.e., low or only moderate JOL accuracy (or gamma correlation) in the immediate-JOL condition arises from something other than the same underlying variable (e.g., memory strength) of JOLs and recall. However, because the focus of Jang and Nelson's study was on the cue-utilization framework and its predicted differences between intrinsic versus extrinsic cues, the reported

state-trace analysis was performed separately for immediate and delayed JOLs. In the currently reported reanalysis, we ascertain whether the monotonic result holds when including the manipulation of immediate versus delayed JOLs in the same state-trace plot (i.e., state-trace plots across eight coupled data points, rather than four coupled data points analyzed separately for immediate and delayed JOLs).

A multidimensional model can produce a monotonic function

An important caveat to the results of Jang and Nelson (2005) is the realization that a monotonic state-trace result does not rule out the possibility of a multidimensional model. Instead, a monotonic result is fully compatible with a multidimensional model if the multidimensional structure happens to project onto the observed 2D subspace in just the right way. For example, consider a function that looks like the letter W, which is clearly nonmonotonic and would falsify a single-dimensional model if viewed from one perspective. However, if the W is a 3D structure (e.g., block lettering) and the perspective on the W is from above, it will appear as a straight line, mistakenly suggesting a single-dimensional structure. Thus, Jang and Nelson may have failed to find the circumstances necessarily to reveal the multidimensional structure: it may be that a different choice of dependent or independent variables would produce a multidimensional state-trace result (possibly consistent with the cue-utilization framework, but perhaps a multidimensional model of a different kind). As explained next, we consider, in turn, the possibility that a different dependent measure (latency) or different manipulations (timing of JOLs or test practice) may produce a multidimensional state-trace result.

Recall accuracy and latency as measures of memory

Although recall accuracy has been mainly used as a measure of memory, there have been arguments that a different measure is more appropriate under certain conditions. For example, the use of dichotomous criterial measures, such as accuracy, may yield qualitatively misleading results owing to insensitivity of the measurement (Loftus, 1978, 1985; MacLeod & Nelson, 1984). By contrast, recall latency may continue to find differences between two conditions even when accuracy is the same for both conditions (Wearing & Montague, 1970). As applied to the current situation, consider the role of item variability, such that for a given participant, some words on a list stand out and are easily recalled whereas others are difficult to remember. In this case, studying the difficult items for an even longer duration might fail to make them recallable. At the same time, studying the easy items for a longer duration might result in stronger memories for those items. However, this strengthening of the easy items might not be apparent when using accuracy as the dependent measure considering that these items would have been recalled regardless of study of duration (i.e., they were already above criterion). Nevertheless, if these easy items have been strengthened, this might be revealed through recall latency (i.e., the longer study duration produces faster recall, albeit faster recall of the same subset of items that would have been recalled with a shorter study duration). This is a hypothetical example, and more generally, study duration is likely to affect recall accuracy. However, this example makes the point that even in situations where accuracy is not at ceiling or floor, the criterial nature of recall accuracy indicates that changes in memory strength may be revealed as changes in recall latency even if there is no change in recall accuracy.

Beyond the possibility that recall latency may be a more sensitive measure of memory strength, consider the possibility that memory structures are multidimensional, with different dependent measures (e.g., recall accuracy versus latency) tapping different aspects of a memory (e.g., Millward, 1964). For instance, MacLeod and Nelson (1984) proposed that "error probability measures the sufficiency of the encoding for retrieval, whereas correct latency measures the number of decoding steps during retrieval before the item is output" (pp. 233–234). They reached this conclusion after observing that test practice, as compared to study practice, produced opposite effects on recall accuracy and latency, producing less accurate recall that was nevertheless faster. Thus, if some of the manipulations used by Jang and Nelson (2005) affected the latter aspect of memory (i.e., the number of decoding steps needed for recall) whereas JOLs primarily reflected the former aspect of memory (i.e., the sufficiency of encoding), then a multidimensional state-trace result may be apparent between JOLs and memory performance when the measure of performance is recall latency rather than accuracy. Indeed, the recently proposed Primary and Convergent Retrieval memory model (Hopper & Huber, 2018), assumes that some forms of learning (e.g., rote rehearsal) strengthen associations between retrieval cues and the target memory, supporting the sufficiency of memory, whereas other forms of learning (e.g., retrieval practice) strengthen associations between the features that comprise a target memory, resulting in fewer decoding steps and faster retrieval. If participants engage in covert retrieval practice when making a JOL, and such covert retrieval is more effective in one condition than in another (e.g., delayed- versus immediate-JOL conditions: Spellman & Bjork, 1992), then an examination of latency may reveal a previously hidden dissociation between JOLs and recall performance. This possibility is examined with a new analysis of Jang and Nelson's data (and those of a new experiment, which will be described later) by including latency in the analysis.

Mnemonic cues: Beyond intrinsic and extrinsic cues

Consider manipulations designed to selectively affect JOLs, in contrast to intrinsic and extrinsic cues, which affected both JOL magnitude and recall accuracy in the experiments reported by Jang and Nelson (2005). Ironically, this possibility is suggested by the cue-utilization framework. Beyond intrinsic versus extrinsic cues, Koriat (1997) proposed mnemonic cues as a third class of cues, which are internal "indicators that may signal for the participant the extent to which an item has been learned and will be recalled in the future" (p. 351). In other words, this third class of cues is mostly related to the metacognitive awareness that the participant may have regarding a particular target item. According to the cue-utilization framework, intrinsic and extrinsic cues affect JOLs not only directly but also indirectly through their influences on internal, mnemonic cues. Critically, the direct and indirectly mediated effects are assumed to entail qualitatively different processes. The direct effects of intrinsic and extrinsic cues are likely to involve an analytic inference based on the person's a priori theory about the memory-related consequences: e.g., "I should be able to recall this item because I have good memory for names (intrinsic cues) or because I studied it for longer (extrinsic cues)". The effects of mnemonic cues, in contrast, involve a nonanalytic, implicit inference, rather than logical deduction: e.g., "I just know that I will be able to recall this item on the later test".

It is not entirely clear how to manipulate mnemonic cues independent of intrinsic and extrinsic cues in the context of the cue-utilization framework because mnemonic cues are assumed as mediators, which are sensitive to both intrinsic and extrinsic factors. Nonetheless, one possibility is the manipulation of immediate versus delayed JOLs, and this is addressed by including the two JOL conditions in the state-trace analysis.

Another manipulation that might selectively boost mnemonic cues is practice (although practice can affect recall as well as

JOLs), or more precisely, prior test experience. According to Koriat (1997), "the relative weight of different cues in determining JOLs may differ from one condition to another and may also change with practice studying the same list of items. …the increased reliance on mnemonic cues with practice may be expected to improve JOL accuracy because such cues reflect the effects of past experience and can serve as a good basis for memory predictions" (pp. 351–352). Supporting this claim, Koriat had participants cycle through the study (S), JOL rating (J), and test phases (T) more than once, and JOL accuracy increased as a function of study-JOL-test (SJT) cycle (see also, Koriat, Sheffer, & Ma'ayan, 2002): e.g., higher gamma correlations on the second SJT cycle than on the first SJT cycle. Prior test practice has the potential to improve overall memory performance, particularly with a delayed final test (see Roediger & Karpicke, 2006, for a review). However, in the absence of feedback during test practice, and with an immediate final test (e.g., less than 5 min after practice), test practice is unlikely to increase recall accuracy as compared to the control condition (e.g., Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012). Yet, test experience will likely affect item-specific mnemonic cues that determine JOLs: i.e., prior test experience recalling items makes it easier to know which items are recallable.

From Koriat and colleagues (Koriat, 1997; Koriat et al., 2002), it is not clear whether the JOL accuracy benefits of prior test experience uniquely reflect test practice (T) or whether they arise from additional study (S) and/or prior JOL (J) experience (their design involved a full cycle through SJT before the final SJT cycle). However, other results suggest that the benefits uniquely reflect test experience (e.g., Jang, Wallsten et al., 2012; King, Zechmeister, & Shaughnessy, 1980; Lovelace, 1984). For example, Jang, Wallsten et al. (2012) compared different conditions with prior S, SJ, ST, or SJT on a first cycle through the items before a final SJT cycle. They found that JOL accuracy was improved in ST and SJT conditions, relative to S and SJ conditions for immediate JOLs (and the delayed-JOL effect was obtained regardless of the practice conditions). This JOL accuracy benefit of prior test experience in the immediate-JOL condition is termed the testing-JOL effect. These findings suggest that on a first test, different latent variables are used for immediate JOLs whereas after prior test practice, mnemonic cues are used for JOLs, producing a closer correspondence between immediate JOLs and recall performance.

New experiment and method

In summary, Jang and Nelson (2005) found consistent support for the sufficiency of a single dimension for explaining JOL magnitude and recall accuracy (this was true for separate analyses of immediate and delayed JOLs) across a variety of intrinsic and extrinsic cues. However, the delayed- and testing-JOL effects indicate that in some circumstances, JOLs rely on information that is different from the information that underlies recall, resulting in a low gamma correlation (or low JOL accuracy) whereas JOLs are highly predictive after a delay or if there is prior test experience with the study items, resulting in a high gamma correlation (or high JOL accuracy). These effects suggest a multidimensional function, with immediate JOLs in the absence of prior test experience being affected by factors other than those responsible for recall performance. Thus, it may be that Jang and Nelson simply failed to consider dependent variables or manipulations necessary to reveal this multidimensional function. In the present study, we report a reanalysis of Jang and Nelson's data that addresses both recall accuracy and latency, including both immediate- and delayed-JOL conditions in each state-trace plot. In addition, we report the data of a new experiment in which test experience was manipulated through SJT cycles while manipulating intrinsic and extrinsic cues.

We briefly describe the new experiment that yielded data to which we applied state-trace analysis, followed by the methodology of state-trace analysis for each of the six experiments (the first five from Jang & Nelson, 2005). Although each of the Jang and Nelson's experiments included a manipulation of immediate and delayed JOLs (i.e., for the delayed-JOL effect), as well as intrinsic and extrinsic cues, none of them included the manipulation of test practice (no test versus prior test). In the new experiment, we collected the data for the testing-JOL effect, including an intrinsic cue and an extrinsic cue as in Jang and Nelson's experiments. Specifically, the procedure of the experiment was similar to that of Jang and Nelson's Experiment 1B in which item relatedness and number of presentations were used as intrinsic and extrinsic cues, respectively, for each of the immediate and delayed JOLs, except that in the new experiment, an initial cued recall test was included before the second study (as summarized in Table 1). In that way, the experiment allowed us to test whether the presence versus absence of prior test experience would yield a nonmonotonic function between JOLs and recall, particularly in the immediate-JOL condition (as gamma correlations for delayed JOLs were high regardless of the manipulation of prior test experience). The materials were the same as those of Experiment 1B. Of 48 word pairs, 24 consisted of nouns that were moderately related, and the remaining 24 consisted of nouns that were not obviously related. One half of the pairs in each condition received immediate JOLs, and the other half received delayed JOLs. During the study phase (S), each pair was presented for 5 s. Participants ($N = 49$) were instructed to study word pairs and to give a JOL (i.e., predict future recall probability of the target: 0%, 20%, 40%, 60%, 80%, and 100%) whenever a cue word appeared alone. Immediate JOLs were elicited immediately after the offset of each pair presented for study, and delayed JOLs were elicited after all the pairs had been studied. Both JOLs (J) and cued recall (T) were self-paced. Once participants finished the first SJT cycle, the second SJT cycle began, and the order of presentation of each pair was randomized anew for study, JOLs and cued recall. No feedback was given.

Because the results reported by Jang and Nelson (2005) revealed overlapping functions that were perfectly monotonic (given four coupled data points of JOLs and recall from two levels of an intrinsic cue and two levels of an extrinsic cue), there was no need to use any statistical tests. However, in the present study, statistical tests are needed to assess the dimensionality of the results when including the manipulation of immediate versus delayed JOLs in the state-trace plots (eight coupled data points for each experiment), considering that the possibility of a false positive nonmonotonic conclusion increases as a function of the number of conditions. In other words, there are more ways in which one condition might deviate from a monotonic function owing to chance alone when there are more conditions. Therefore, we used coupled monotonic regression (CMR; Kalish, Dunn, Burdakov, & Sysoev, 2016; see also, Dunn et al., 2012; Newell et al., 2010) to determine the statistical reliability of the conclusion as regards dimensionality. Although the CMR procedure could be applied to three-dimensional data, we apply it to pairs of two dependent measures. We do so for simplicity (e.g., 2D plots are easier to interpret) and to make contact with the prior analyses, but also because the finding of a nonmonotonic function (as is the case in the reported analyses) for any pair of dependent measures is sufficient for ruling out the default single-dimensional model. The CMR procedure consists of two parts: model fitting and then model testing. The model fitting part yields the fit of an order-restricted two-dimensional model (nonmonotonic state trace) and the fit of an order-restricted single-dimensional model (monotonic state trace) to the observed data. The model testing part compares goodness of fit for the two models and produces a difference in goodness of fit (ΔG^2). Then, the empirical distribution of ΔG^2 is estimated using a Monte Carlo

Table 2
CMR (Coupled Monotonic Regression) test results.

JOL condition	Experiment	JOL magnitude; Recall latency		Recall latency; Accuracy		Recall accuracy; JOL magnitude	
		ΔG^2	p	ΔG^2	p	ΔG^2	p
Both	1A	**11.27**	**.005**	1.68	.396	0	1
	1B	2.35	.274	2.78	.227	0.06	.823
	1C	**49.94**	**<.001**	**3.81**	**.097**	**14.45**	**.001**
	1D	**16.01**	**.003**	**11.07**	**.018**	0.05	.783
	2	2.49	.288	**7.43**	**.030**	0.03	.898
	New	**9.38**	**.016**	**6.57**	**.059**	0.02	.898
Immediate	1A	0	.993	0	.999	0	.999
	1B	0	.973	0	.997	0	.999
	1C	0	.856	0	.893	0	.954
	1D	0	.949	0	.975	0	.972
	2	0.09	.304	0.09	.308	0	.992
	New	**6.87**	**.004**	0.09	.310	0	1
Delayed	1A	0	.999	0	1	0	1
	1B	0.03	.320	0.03	.326	0	.997
	1C	0	.944	0	.956	0	.951
	1D	0	.998	0	.997	0	.999
	2	0	.995	0	.998	0	1
	New	0	1	0.001	.378	0.001	.408

Note. Bold fonts indicate a statistically significant rejection of the single-dimensional model with an alpha of .05 or marginal significance ($.05 < p < .10$). JOL = Judgment of learning.

simulation with a large number of samples (e.g., 10,000 used in the study).

For each data set of the six experiments, two order constraints were applied to rule out nonsensical orderings of the conditions. First, it was assumed that performance should not increase from the condition of related or easy pairs to the condition of unrelated or difficult pairs (intrinsic cues). Second, it was assumed that performance should not increase from the condition of two presentations or long duration to the condition of one presentation or short duration (extrinsic cues). A series of state-trace analyses, using CMR, was conducted for the dependent measures of (1) JOL magnitude and recall latency; (2) recall accuracy and latency; and (3) JOL magnitude and recall accuracy. In addition, the analysis was conducted in the three different ways: (i) including the manipulation of immediate versus delayed JOLs; (ii) only for immediate JOLs; and (iii) only for delayed JOLs. Thus, the state-trace analysis of JOLs and recall accuracy as reported in Jang and Nelson (2005) are here duplicated in part (ii and iii), but with CMR to assess statistical reliability, and the analysis is extended to include the manipulation of immediate versus delayed JOLs, the dependent variable of latency, and a new testing-JOL experiment. It should be noted that for the data of the new experiment, state-trace analysis was applied completely in the three ways (i to iii) because it was also necessary to examine whether we would replicate the findings of Jang and Nelson: i.e., monotonic functions between immediate JOLs and recall accuracy (ii); and between delayed JOLs and recall accuracy (iii). Individual trial recall latencies for correctly recalled items in each condition were transformed using the natural logarithm, prior to statistical analyses. Furthermore, because the CMR algorithm is based on an assumption of a monotonically increasing function, rather than a monotonically decreasing function, log latencies were subtracted from a constant prior to state-trace analyses.

Before turning to the results, we consider a couple of concerns that arise with the use of latencies: whether differences between conditions might arise from either list-composition effects or response caution. Free recall latencies increase as a function of the number items recalled so far within the test period, indicating sampling competition between items from the same list (e.g., Rohrer & Wixted, 1994). More specifically, with a mixed list of items that differ in strength, strong items are recalled earlier and more quickly (Wixted, Ghadisha, & Vera, 1997). However, these list-composition effects are found with free recall but not cued recall; because the cue focuses the memory search on the target memory, list composition plays little or no role for cued recall. In support of this claim, a recent study with 5 experiments failed to find list-strength effects with cued recall even though list-strength effect were readily found with free recall (Wilson & Criss, 2017). Besides list composition, another concern with latencies is the possibility of a speed–accuracy tradeoff, which could potentially produce relatively uninteresting latency effects (i.e., latency effects that reflect a change in response caution, rather than a change in the properties of the to-be-recalled memories). However, differences in speed–accuracy tradeoff between conditions are highly unlikely in the current situation. Specifically, with the currently employed mixed-list design, such differences would require that the participant is aware of the condition associated with the cue word on each test trial, using this information to adopt a different level of response caution in their recall attempt (e.g., how willing they are to make a guess). Nevertheless, because the cue words were drawn from the same pool of words for all conditions, the only way that the participant could be aware of the associated condition would be to explicitly recall the study circumstances, in which case, they are likely to have already recalled the target word.

Results

Table 2 shows the results of the CMR test which was conducted for the dependent measures of (1) JOL magnitude and recall latency; (2) recall accuracy and latency; and (3) JOL magnitude and recall accuracy, (i) when both JOL conditions were included for the simultaneous fits of immediate and delayed JOLs; and (ii and iii) when one of the two conditions was included for the separate fits.

State-trace plots of JOLs and recall latency: Delayed-JOL effect (all experiments)

Fig. 2 shows the outcome of the state-trace analysis of JOL magnitude and recall latency for each experiment where each plot includes both immediate and delayed JOLs (simultaneous fits). In each panel of the figure, there are four coupled data points for each of the two JOL conditions (circles for the immediate-JOL condition, and triangles for the delayed-JOL condition), which represent two levels of an intrinsic cue and two levels of an extrinsic cue. In

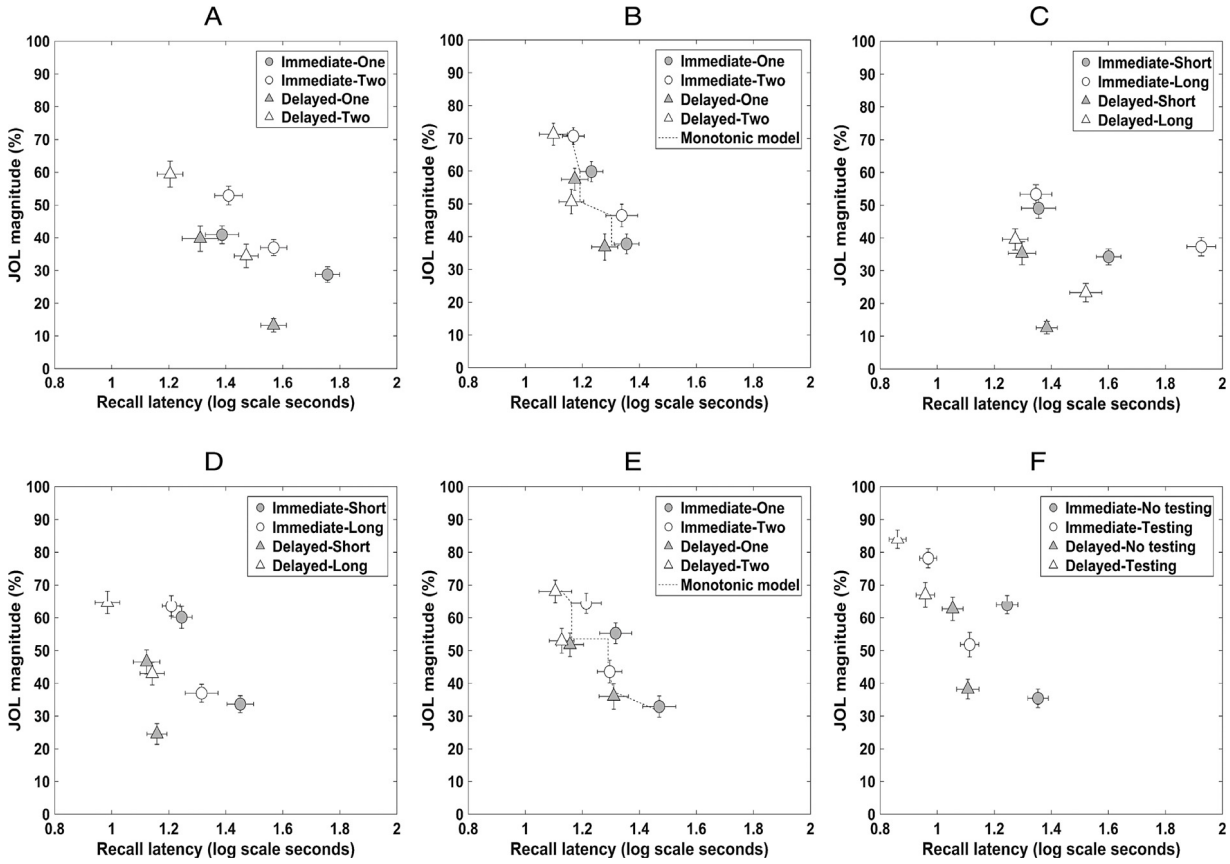

Fig. 2. State-trace plots of recall latency and JOL (judgment of learning) magnitude for each of the five experiments, 1A to 1D, and 2 (Jang & Nelson, 2005); and the new experiment: Panels A to D = Experiments 1A to 1D, respectively; Panel E = Experiment 2; and Panel F = new experiment. For each experiment, a 2 (timing of JOLs) × 2 (intrinsic cue) × 2 (extrinsic cue) repeated-measures design was used: for intrinsic cues, Panels A and C = item difficulty; Panels B and D to F = item relatedness: for extrinsic cues, Panels A, B, and E = number of presentations; Panels C and D = study duration; Panel F = first SJT (no testing) versus second SJT (testing). For each panel, filled and unfilled circles = low and high levels of the extrinsic cue in the immediate-JOL condition, respectively; filled and unfilled triangles = low and high levels of the extrinsic cue in the delayed-JOL condition, respectively. For each of the four symbols, two data points correspond to the two levels of the intrinsic cue. Each vertical and horizontal hash mark depicts the standard error of the mean. Dashed lines (Panels B and E) indicate the best-fitting monotonic model, which was not rejected. Each panel with no dashed lines shows that the monotonic model was rejected.

the event that the single-dimensional model was not rejected, the best-fitting monotonic function is illustrated by dashed lines. The single-dimensional model was rejected in each of the four experiments (Experiments 1A, 1C, and 1D; and the new experiment), as illustrated in Panels A, C, D, and F, respectively: $\Delta G^2 s > 9.38$, $ps \leq .016$. However, we were unable to reject the single-dimensional model in Experiments 1B and 2, as illustrated in Panels B and E, respectively: $\Delta G^2 s < 2.49$, $ps \geq .274$. In general, the results suggest that extra flexibility of a multidimensional model is needed in most circumstances to explain the relationship between JOLs and recall latency when considering both immediate and delayed JOLs.

State-trace plots of JOLs and recall latency: Testing-JOL effect (new experiment)

Panel F of Fig. 2 shows the results from the new experiment, which examined the testing-JOL effect (first versus second cycle of SJT) as well as the delayed-JOL effect. To reveal the unique contribution of prior test experience, apart from any nonmonotonicity arising from the manipulation of immediate versus delayed JOLs, we report separate analyses for immediate and delayed JOLs. The four coupled data points of the immediate-JOL condition are shown in Panel A of Fig. 3, and the four coupled data points of the delayed-JOL condition are shown in Panel B. The state trace for immediate JOLs is indeed nonmonotonic: $\Delta G^2 = 6.87$, $p = .004$ (which requires a multidimensional model) whereas the state trace for delayed JOLs is monotonic: $\Delta G^2 = 0$, $p = 1$.

One lingering question is whether this nonmonotonic function for immediate JOLs and recall latency reflects the fact that immediate JOLs were used, or whether the manipulation of prior test experience was included. To investigate this possibility, we additionally conducted state-trace analyses for immediate JOLs and recall latency of Jang and Nelson's experiments, which did not include prior test experience. In contradiction to the new experiment for the testing-JOL effect, these state-trace analyses revealed that the single-dimensional model was sufficient for immediate JOLs and recall latency: $\Delta G^2 s < 0.09$, $ps \geq .304$, for all five of Jang and Nelson's experiments (this was equally true for delayed JOLs and recall latency: $\Delta G^2 s < 0.03$, $ps \geq .320$, for all five of Jang and Nelson's experiments). These findings indicate that the departure from monotonicity observed in the immediate-JOL condition of the new experiment is due to the inclusion of prior test practice.

State-trace plots of recall accuracy and latency (all experiments)

When including both JOL conditions and analyzing memory performance in terms of latency, the state-trace results were mostly nonmonotonic (as shown in Fig. 2). However, it is not immediately clear whether these apparent nonmonotonic functions reflect dissociations between JOLs versus memory performance

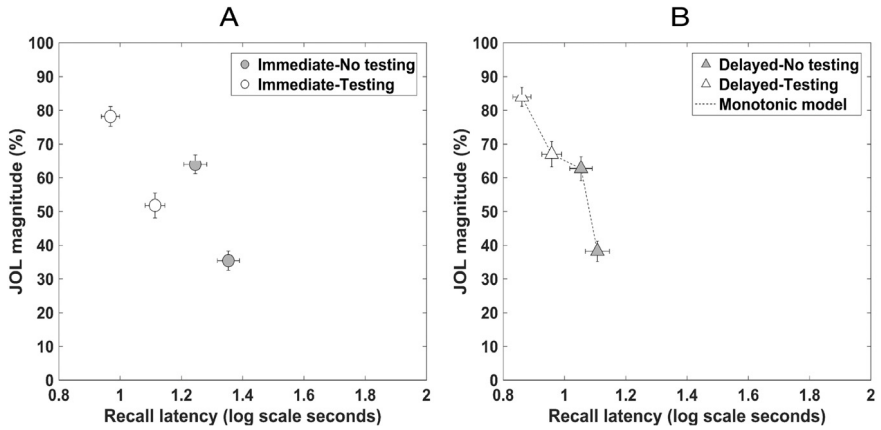

Fig. 3. State-trace plots of recall latency and JOL (judgment of learning) magnitude for the immediate-JOL condition (Panel A) and delayed-JOL condition (Panel B) of the new experiment: The data of Panel F in Fig. 1 split into these two data sets. A 2 (immediate versus delayed JOLs) × 2 (unrelated versus related pairs) × 2 (first SJT [no testing] versus second SJT [testing]) repeated-measures design was used: filled circles (Panel A) and filled triangles (Panel B) = unrelated and related pairs in the no-testing condition; unfilled circles (Panel A) and unfilled triangles (Panel B) = unrelated and related pairs in the testing condition. Each vertical and horizontal hash mark depicts the standard error of the mean. Panel A shows that the monotonic model was rejected while Panel B shows the monotonic model (as illustrated by dashed lines) was not rejected.

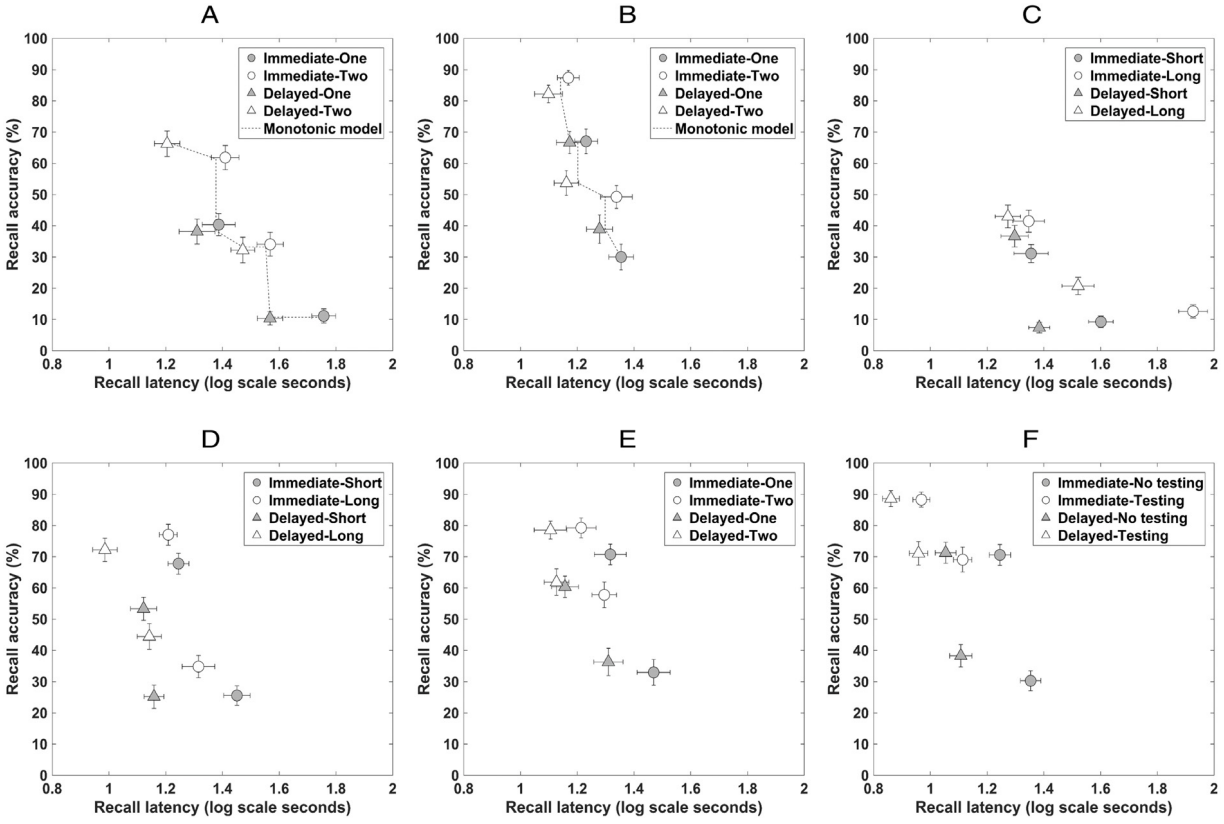

Fig. 4. State-trace plots of recall latency and accuracy for each of the five experiments, 1A to 1D, and 2 (Jang & Nelson, 2005); and the new experiment: Panels A to D = Experiments 1A to 1D, respectively; Panel E = Experiment 2; and Panel F = new experiment. For each experiment, a 2 (timing of JOLs) × 2 (intrinsic cue) × 2 (extrinsic cue) repeated-measures design was used: for intrinsic cues, Panels A and C = item difficulty; Panels B and D to F = item relatedness: for extrinsic cues, Panels A, B, and E = number of presentations; Panels C and D = study duration; Panel F = first SJT (no testing) versus second SJT (testing). For each panel, filled and unfilled circles = low and high levels of the extrinsic cue in the immediate-JOL condition, respectively; filled and unfilled triangles = low and high levels of the extrinsic cue in the delayed-JOL condition, respectively. For each of the four symbols, two data points correspond to the two levels of the intrinsic cue. Each vertical and horizontal hash mark depicts the standard error of the mean. Dashed lines (Panels A and B) indicate the best-fitting monotonic model, which was not rejected. Each panel with no dashed lines shows that the monotonic model was rejected.

more broadly, or whether they reflect dissociations between recall accuracy and latency (with recall accuracy tracking with JOLs whereas latency reflects some other processes). To assess this situation, we conducted state-trace analysis between the two measures of memory performance (recall accuracy versus latency) for each experiment when both immediate- and delayed-JOL conditions were included. The monotonic function was rejected for each of the four experiments (Experiments 1C, 1D, and 2; and the new experiment), as illustrated in Panels C, D, E, and F of Fig. 4, respectively: $\Delta G^2 s < 3.81$, $ps \leq .097$. However, the single-dimensional

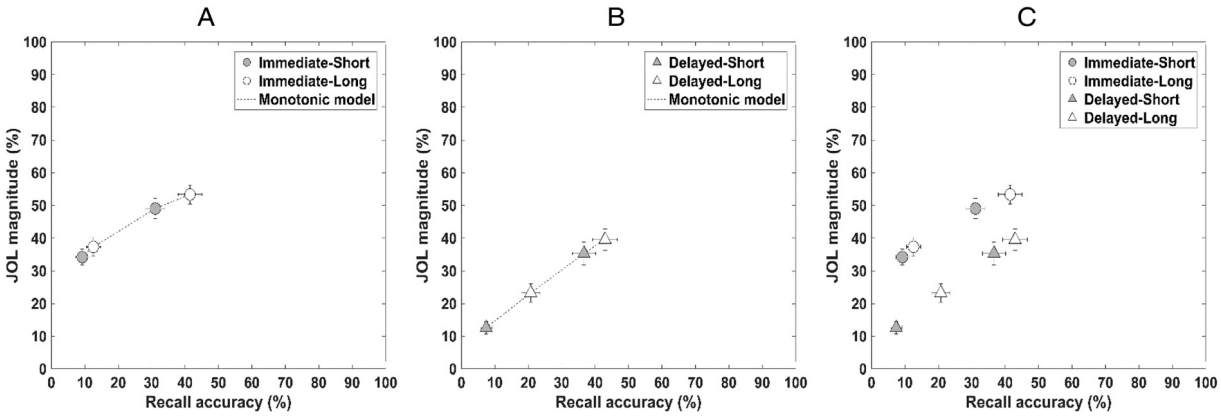

Fig. 5. State-trace plots of recall accuracy and JOL (judgment of learning) magnitude of Experiment 1C (Jang & Nelson, 2005): Panels A and B show the data of the immediate condition and delayed condition, respectively, and Panel C shows those of both conditions. A 2 (immediate versus delayed JOLs) × 2 (difficult versus easy pairs) × 2 (short versus long study duration) repeated-measures design was used: filled and unfilled circles (Panels A and C) = short and long durations in the immediate-JOL condition, respectively; filled and unfilled triangles (Panels B and C) = short and long durations in the delayed-JOL condition, respectively. For each of the four symbols, two data points correspond to the two levels of difficult and easy pairs. Each vertical and horizontal hash mark depicts the standard error of the mean. Each of the panels, A and B, shows the monotonic model (as illustrated by dashed lines for each) was not rejected, but Panel C shows that the monotonic model was rejected.
Source: Panels A and B: adapted from Fig. 6 in "How many dimensions underlie judgments of learning and recall? Evidence from state-trace methodology", by Y. Jang, and T. Nelson, 2005, *Journal of Experimental Psychology: General, 134*, p. 316.

model was not rejected in Experiments 1A and 1B, as illustrated in Panels A and B, respectively, with dashed lines: $\Delta G^2 s > 2.78$, $ps \geq .227$. These results indicate that there is indeed a dissociation between recall accuracy and latency, which may underlie the observed dissociations between JOLs and recall latency in some cases (e.g., Experiments 1C and 1D; and the new experiment). However, this dissociation between recall latency and accuracy does not fully explain the JOL results. First, the state trace of Experiment 1A between recall accuracy and latency (Panel A of Fig. 4) supports a single-dimensional account of memory whereas this experiment revealed a multidimensional account of JOLs and recall latency (Panel A of Fig. 3). Second, the state trace of Experiment 2 supports a multidimensional account of recall accuracy and latency (Panel E of Fig. 4) but a single-dimensional account of JOLs and recall latency (Panel E of Fig. 3).

We further conducted state-trace analysis between recall accuracy and latency for each of the immediate- and delayed-JOL conditions, separately. The results revealed that the single-dimensional model was not rejected in all 12 data sets (i.e., six experiments × two JOL conditions): $\Delta G^2 s < 0.09$, $ps \geq .308$.

State-trace plots of JOLs and recall accuracy (all experiments)

Finally, state-trace analysis of JOL magnitude and recall accuracy was conducted for each experiment when including the manipulation of immediate versus delayed JOLs. The state trace across the eight coupled data points supported the single-dimensional model in every case, $\Delta G^2 s < 0.06$, $ps \geq .783$, except for Experiment 1C[1] : $\Delta G^2 = 14.45$, $p = .001$. First, the monotonic function found in each of the five experiments (Experiments 1A, 1B, 1D, and 2; and the new experiment, as well) was consistent with perfect monotonicity of Jang and Nelson (2005) in which state-trace analysis was performed separately for each of the immediate- and delayed-JOL conditions (i.e., with four coupled data points). Specifically, for immediate JOLs, the model fit (ΔG^2) was approximately zero for all six experiments, including the new experiment ($ps \geq .954$). For delayed JOLs, this was equally true not only for all five of Jang and Nelson's experiments ($ps \geq .951$) but also for the new experiment ($p = .408$). These findings indicate that the support for the single-dimensional account of JOLs and recall reported by Jang and Nelson was not an artifact of failing to analyze the data for both JOL conditions simultaneously. Second, the sole exception is presented in Fig. 5, which shows the state-trace plots of JOLs and recall accuracy for Experiment 1C, revealing a nonmonotonic function when the immediate and delayed conditions are put into the same state-trace plot (Panel C) even though these functions are monotonic when the immediate (Panel A) and delayed (Panel B) JOL conditions are analyzed separately.

In general, these state-trace plots comparing JOL magnitude and recall accuracy support Jang and Nelson's (2005) conclusions regarding the intrinsic and extrinsic cues from the cue-utilization framework (Koriat, 1997). That is, according to the cue-utilization framework, JOLs and recall accuracy should have been differently affected for these particular experimental manipulations. However, monotonic state-trace results are only useful if both of the dependent measures are sufficiently affected by the manipulations. Specifically, a multidimensional model will necessarily produce a monotonic state-trace plot if one of the dependent measures is unchanged by the manipulations, such as could occur at ceiling of floor, or if the manipulations are too weak. Jang and Nelson addressed this issue with an analysis of variance (ANOVA) of JOLs and recall accuracy separately, and in the Appendix, we report the corresponding ANOVAs for the new experiment. Of note, such a manipulation check is not necessary for the recall latency results considering that the state-trace plots were clearly nonmonotonic when recall latency was compared to either JOL magnitude or recall accuracy; if recall latency had not been affected by the experimental manipulations, those state-trace plots would have been monotonic.

Discussion

Using state-trace plots, we reanalyzed the data of Jang and Nelson (2005) and a new experiment to revisit the conclusion that a single latent variable underlies both JOLs and recall. Jang and Nelson tested Koriat's (1997) cue-utilization framework, which predicted that intrinsic cues should affect JOLs and recall equivalently whereas extrinsic cues should affect recall more than JOLs, indicating a two-dimensional model. In contradiction to the prediction from the cue-utilization framework, all 10 state-trace plots from the five experiments of Jang and Nelson (five for immediate JOLs and the remaining five for delayed JOLs), revealed monotonic

[1] We are deeply indebted to John Dunn for bringing this to our attention.

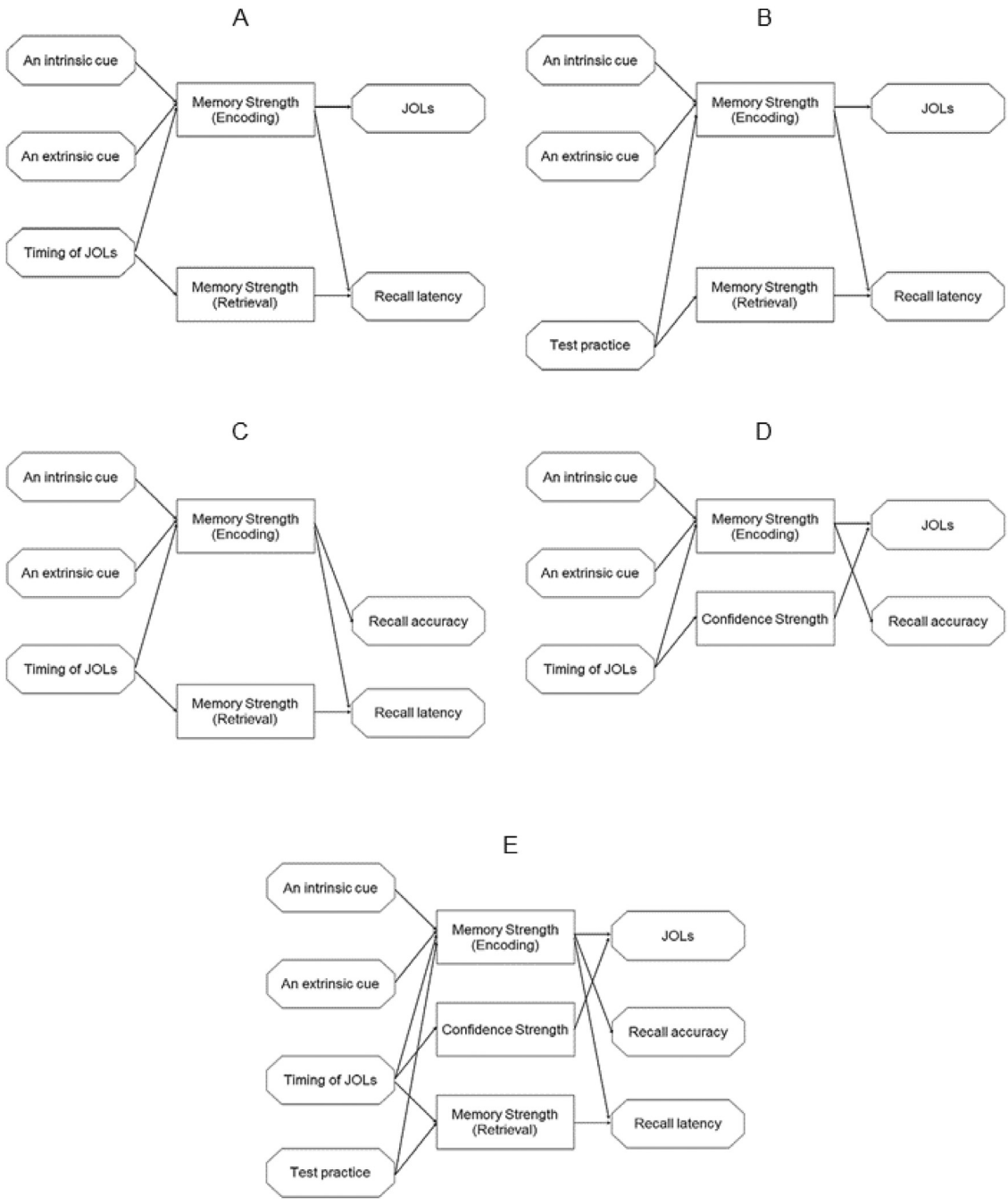

Fig. 6. Two-dimensional models to account for the results of Jang and Nelson's (2005) experiments and the new experiment (Panels A to D); and a three-dimensional model that incorporates the two-dimensional models (Panel E). See text for explanation.

functions, suggesting that a single latent variable (e.g., memory strength) explains JOLs and memory performance. This conclusion also held for separate analysis of the immediate- and delayed-JOL conditions in the new experiment (i.e., in total, all 12 state-trace plots of monotonic functions). However, we note that the single-dimensional model is nested under a multidimensional model and the reported monotonic functions do not necessarily falsify a multidimensional model. Instead, it may be that the true multidimensional model happened to project onto the 2D subspace defined by the chosen dependent measures and manipulations in such a way as to produce a monotonic function. Our new analyses produced nonmonotonic functions (rejecting the single-dimensional model) when we used recall latency as the dependent measure of memory performance rather than recall accuracy. Furthermore, this was the case for the originally reported data as well as the data of the new experiment that included manipulations sensitive to JOL accuracy (i.e., timing of JOLs and test practice). Additional analysis revealed that this dissociation between JOL magnitude and recall latency was partially, but not fully, explained by dissociations between recall accuracy and latency.

Reiterating an important point made in the Introduction, when considering that these experiments used a self-paced cued recall (rather than free recall) test, and when considering that these experiments used a mixed design in which the participant could not know the experimental condition of the cue word on a test trial prior to successful recall, it is unlikely that these latency differences between conditions reflect changes in the decision process, such as response caution. Rather, differences in retrieval latency between

the conditions likely reflect differences in the magnitude or quality of the target memory.

These results are generally supportive of theories of JOLs, which appeal to factors other than memory strength when explaining improved JOL accuracy (or gamma correlation) in one condition relative to another condition. Although state-trace analysis can be used to identify the number of latent variables underlying the observed dependent measures, it does not identify the nature of these latent variables, which might or might not be consistent with existing theories. Next, we consider several possibilities. Fig. 6 shows candidate two-dimensional models (Panels A to D) for explaining different pairs of dependent measures as well as a three-dimensional model as applied to all three dependent measures (Panel E). For each panel, experimental manipulations included in the analysis, possible latent variables, and dependent measures are illustrated from left to right through arrows. In this way, we consider how latent variables underlie JOLs and recall might explain all results.

A two-dimensional model with two different aspects of memory strength

The first three panels of Fig. 6 (Panels A to C) present a specific two-dimensional model which assumes that there are two different aspects of memory strength, with one related to the sufficiency of encoding and the other unique to the retrieval process. The encoding strength latent variable is assumed to affect JOLs, recall accuracy, and recall latency whereas the retrieval strength latent variable uniquely affects latency. Specifically, Panel A (for the delayed-JOL effect) captures the nonmonotonic state-trace plots between JOL magnitude and recall latency observed in four experiments (Experiments 1A, 1C, and 1D; and the new experiment), in which recall latency was shorter in the delayed-JOL condition than in the immediate-JOL condition. Panel B (for the testing-JOL effect) captures the nonmonotonic state-trace plot between JOL magnitude and recall latency observed in the immediate-JOL condition of the new experiment, in which recall latency was shorter after prior test experience. Panel C captures the nonmonotonic state-trace plots between recall accuracy and latency observed in four experiments (Experiments 1C, 1D, and 2; and the new experiment), in which recall latency was shorter in the delayed-JOL condition than in the immediate-JOL condition. This model suggests that covert retrieval practice (such as might occur more effectively when giving a JOL rating after a delay) as well as overt retrieval practice (such as occurs with prior testing experience) leads to faster recall. To the best of our knowledge, this is the first report of faster recall in the delayed-JOL condition as compared to the immediate-JOL condition, despite no difference in recall accuracy between the two conditions (although JOL studies do not typically consider recall latency).

The concept of retrieval strength as a second aspect of memory is consistent with the conclusion that study practice has a larger effect on encoding but a smaller effect on retrieval, as compared to test practice (Birnbaum & Eichner, 1971; Hogan & Kintsch, 1971). For example, MacLeod and Nelson (1984) found shorter recall latency but lower recall accuracy immediately after four testing cycles in comparison to three study cycles and one testing cycle (i.e., STTT versus SSST: see also, van den Broek, Segers, Takashima, & Verhoeven, 2014). Similarly, an immediate final test following test practice without feedback often fails to increase recall accuracy even relative to the control (with no practice) condition (e.g., Jang, Wixted et al., 2012) although a hidden benefit of this test practice is revealed by analyzing recall latency (e.g., Hopper & Huber, 2018).

This model explains the nonmonotonic function between JOL magnitude and recall latency for the new testing-JOL experiment when analyzing the data only for immediate JOLs. In this case, prior test experience increased retrieval strength, producing faster recall, but this increase in retrieval strength had little impact on the on-average JOL magnitude. However, JOL accuracy, as measured by the gamma correlation, increased with prior test experience (i.e., the testing-JOL effect). Thus, this model implies some sort of item-by-item differentiation process to explain why prior test experience produced no or little change in the on-average JOL magnitude across items even though prior test experience produced a better ability to predict which items would be recalled. In other words, it must be that while JOL magnitude increased for items that would be recalled as a function of prior test experience, JOL magnitude decreased for items that would not be recalled as a function of prior test experience. That this nonmonotonicity only occurred with the immediate-JOL condition is sensible when considering the hypothesis that a delayed JOL involves a covert retrieval attempt (or does that more effectively than an immediate JOL), in which case even the no-testing condition has some level of covert retrieval practice experience. The question of whether the delayed-JOL effect reflects covert retrieval is detailed in competing accounts of the delayed-JOL effect (Nelson & Dunlosky, 1991; Spellman & Bjork, 1992; see also, Jang, Wallsten et al., 2012), and the success of this model may shed light on this debate.

A two-dimensional model with memory strength and confidence

The two different aspects of memory strength model presented above can explain most, but not all of the reported dissociations. Specifically, the new analysis of JOL magnitude and recall accuracy that included both immediate and delayed JOLs produced a nonmonotonic function in one case (Experiment 1C). Perhaps, this finding may be an outlier although in light of the high reliability of this conclusion (Panel C of Fig. 5), this deserves additional consideration. Thus, we consider the possibility that there may be a latent variable that is unique to JOLs, as shown in Panel D of Fig. 6.

For this model, we assumed that JOLs (or more generally, metacognitive judgments) are based on not only memory strength but also another kind of information, or 'confidence strength', explaining the monotonic function between JOL magnitude and recall accuracy found in Experiment 1C, such that the immediate-JOL condition boosted JOL magnitude without affecting recall accuracy. This nonmonotonicity is consistent with the finding from Rhodes and Tauber's (2011) meta-analysis, which concluded that participants are overly confident when making an immediate JOL as compared to a delayed JOL. That is, the dissociation between JOL magnitude and recall accuracy may reflect a change in JOL calibration (i.e., a shift in use of the JOL scale) that is concomitant with the change in JOL accuracy (i.e., knowing which items will or will not be recalled in the future). For some reason, this overconfidence effect was more pronounced in Experiment 1C: The same overconfidence pattern was found in all experiments, but only for Experiment 1C, was it of such a magnitude as to produce a nonmonotonic function between JOLs and recall accuracy.

Although it is not clear what causes overconfidence for immediate JOLs, one possibility is suggested by the claim that ease of retrieval can mislead metacognitive judgments (e.g., Benjamin, Bjork, & Schwartz, 1998). It is likely that people monitor information retrieved from both short- and long-term memory (STM and LTM) when making a JOL, and in general, information is retrieved more quickly from STM than from LTM (Wescourt & Atkinson, 1973). STM information retrieved during immediate JOLs is strong (i.e., the just studied cue–target pair is likely still in STM at the time of the immediate JOL), not only producing overconfidence, but also adding noise to the prediction of subsequent recall (Nelson & Dunlosky, 1991) because such information is not available at the time of recall. STM information is absent for delayed JOLs, resulting in less confident JOLs and also allowing relative differences in JOLs to reflect the more diagnostic LTM information.

A three-dimensional model combining both two-dimensional models

We now consider a three-dimensional model that incorporates both of the above-mentioned two-dimensional models, as seen in Panel E of Fig. 6. Particularly, this three-dimensional model is needed to explain the JOL magnitude and recall accuracy results of Experiment 1C as well as the JOL magnitude and recall latency results from that experiment. Experiment 1C used Swahili–English word pairs (difficult versus easy pairs for the intrinsic cue), and so even the easy pairs would be unfamiliar as compared to other experiments, which used English–English word pairs. In addition, this experiment used study duration (short versus long duration for the extrinsic cue), which is a weaker manipulation as compared to other experiments, which used the number of presentations (e.g., Malmberg & Shiffrin, 2005). For this particular combination of factors (e.g., unfamiliar word pairs that were studied only once), it may be that nondiagnostic confidence-strength information played a stronger role when making JOL ratings. In this case, the results confirm predictions of the cue-utilization framework as regards mnemonic cues. Koriat (1997) distinguished between the rule-based influence underlying the direct effects of intrinsic and extrinsic cues versus the heuristic-based influence underlying the internal, mnemonic cues, and one may refer to Koriat's rule- and heuristic-based influences as confidence strength and the strength of memory encoding, respectively. Critically, the failure to find confidence strength effects in the other experiments does not falsify the cue-utilization framework. Instead, across all experiments the pattern of results seems to indicate that while JOLs primarily reflect memory encoding strength (but not memory retrieval strength, which is unique to recall latency), situations with highly impoverished memories (e.g., second language learning with a single study episode), result in JOLs that more strongly reflect the rule-based process (i.e., not enough information regarding the target memory).

Implications for state-trace methodology

Beyond the implications of our results for JOLs and recall, the results of the present study have important implications for state-trace methodology. Specifically, we make the point that monotonic functions do not rule out the possibility that more than one latent variable underlie the observed data. Instead, the conclusion favoring a single latent variable is made based on parsimony, but there is always the possibility of a hidden nonmonotonicity that would be revealed by choosing different manipulations or different dependent measures.

In terms of different manipulations, the results of the present study indicate the advantage of using higher order factorial designs. Specifically, a 2×2 design can be insufficient for adjudicating between a single-dimensional versus multidimensional model (Loftus et al., 2004). As applied to the current case, when JOL magnitude and recall accuracy were analyzed separately in each of the immediate- and delayed-JOL conditions (i.e., two separate 2×2 designs), the results from all 12 data sets revealed monotonic functions. However, in the case of Experiment 1C, when the same data were analyzed while including both immediate and delayed JOLs (e.g., eight coupled data points, given a $2 \times 2 \times 2$ design) a reliable nonmonotonic function was revealed (also see, Biederman & Tsao, 1979). That is, a nonmonotonicity can be hidden in a 2×2 design although the new experiment produced a nonmonotonic function between JOL magnitude and recall latency not only in a $2 \times 2 \times 2$ design, but also in a 2×2 design of the immediate-JOL condition.

In terms of different dependent variables, the results of the present study indicate that monotonic functions when performance is analyzed with one dependent measure (accuracy) can change to nonmonotonic functions when performance is analyzed with a different dependent measure (latency). Of note, detailed analyses of latency distributions demonstrate that latency is a highly complex dependent measure, potentially reflecting as many as seven different latent variables (e.g., Ratcliff & McKoon, 2008). Thus, perhaps it is not surprising that nonmonotonicity is revealed when using latency. However, the extra latent variables contained within sequential sampling models of reaction time data are largely designed to capture decisional aspects of two-choice performance tasks under time pressure. For instance, there may be a bias for one response over the other that differs between conditions, or one condition may result in slow but accurate responding whereas another condition results in fast error-prone responding. However, in the current case, the condition of a test item could not be known without first recalling the circumstances of initial study, and so these decisional factors were unlikely to vary between conditions. Nevertheless, an important caveat when considering the adoption of additional dependent measures in an attempt to reveal a hidden nonmonotonicity is that the chosen dependent measures need to be task-relevant in a meaningful way for the question of interest.

Another caveat to our use of latency is the finding that a dissociation between accuracy and latency explained some of the reported nonmonotonicity results. Thus, while the adoption of a new dependent measure revealed a previously hidden nonmonotonicity, that nonmonotonicity reflected a more nuanced understanding of memory performance (i.e., the factors that affect accuracy versus those that affect latency). However, this does not invalidate the conclusions regarding JOLs but rather paints a more complete picture of the way(s) in which JOLs are related to memory performance, revealing that JOLs are more sensitive to the latent memory variable that underlies recall accuracy and less sensitive to the latent memory variable that underlies recall latency.

Finally, we reiterate that while state-trace analysis can determine how many latent variables underlie the results, it does not identify the nature of those variables. For instance, the model presented in Panel E of Fig. 6 is one possible account of JOLs and recall, but different interpretations for the underlying processes of JOLs and recall might be reached, based on interaction effects from the traditional methodology (see Loftus, 1978; Loftus et al., 2004, for details) or given different experimental manipulations. More generally, specific models could be tested within a state-trace framework using signed difference analysis (Dunn & James, 2003; Stephens, Dunn, & Hayes, 2018), which asks whether the results are qualitatively compatible with a specific model, rather than simply assessing the number of latent variables.

Conclusions

The present study reversed some of the conclusions reached by Jang and Nelson (2005), revealing dissociations between JOLs and recall when recall performance was measured with recall latency, and in one case when JOL magnitude and recall accuracy were assessed across the full factorial set of conditions that included both immediate and delayed JOLs (Jang and Nelson analyzed each JOL condition separately). This reversal of prior conclusions places an important caveat on the use of state-trace methodology, demonstrating the risk of failing to reject the null hypothesis that a single latent variable explains an observed monotonic function. Instead, it may be that a multidimensional model is closer to the truth, with the observed monotonic function arising from the particular choice of dependent measures and experimental manipulations. Yet, it is not clear from these results whether the important dissociation lies between JOLs and recall latency or whether it lies between recall accuracy and latency: it appears that both forms of dissociation played some role.

Beyond the implications for state-trace methodology, the present study informs theories of JOLs and metacognition. Similar to Jang and Nelson (2005), we failed to find support for the specific two-dimensional model that distinguishes between intrinsic and extrinsic cues (Koriat, 1997). However, unlike Jang and Nelson, we found clear evidence against a unidimensional account of the relationship between JOLs and recall performance where recall latency is a key aspect of performance. In brief, our analyses suggest a different two-dimensional model (i.e., not intrinsic and extrinsic cues). Instead, these results suggest that participants engage in covert retrieval when giving delayed JOLs with successful covert retrieval and become more sensitive to the target memory with retrieval practice through an initial test, decreasing recall latency on the final test without changing recall accuracy. In reaching this conclusion, it is important to note that the instructions given to participants ask them to make predictions of future recall success, and participants are likely to take these instructions to indicate judgments of future recall accuracy rather than future recall latency. It remains to be seen whether the dissociation between JOLs and recall latency observed with the data of both JOL conditions will hold if the instructions focus participants on recall latency (e.g., "rate the probability that you will be able to *quickly* recall the target word on a future recall test"). In any case, because recall accuracy and latency are widely accepted measures of memory performance, these findings pose important challenges for the accounts of JOLs and recall, or more generally, different theories of metacognition.

Appendix. Complete 2 × 2 × 2 (timing of JOLs, extrinsic cue, and intrinsic cue) analysis of variance of the new experiment

	Correct recall			JOL magnitude		
	$F(1, 48)$	p	ES	$F(1, 48)$	p	ES
T	3.04	.088		9.14	.004	.160
E	427.92	<.001	.899	116.13	<.001	.708
I	192.34	<.001	.800	190.30	<.001	.799
T × E	2.69	.108		19.54	<.001	.289
T × I	4.01	.051	.077	8.89	.004	.156
E × I	45.48	<.001	.487	5.38	.025	.101
T × E × I	2.30	.136		2.74	.104	

Note. Effect size (ES) is reported only when the F value was significant. JOL = judgment of learning; T = timing of JOLs (immediate versus delayed JOLs); E = extrinsic cue (first SJT [no testing] versus second SJT [testing]); I = intrinsic cue (unrelated versus related pairs).

References

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*, 137–181. http://dx.doi.org/10.1016/0022-2496(79)90016-6.

Benjamin, A., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*, 55–68. http://dx.doi.org/10.1037/0096-3445.127.1.55.

Biederman, I., & Tsao, Y. C. (1979). On processing Chinese ideographs and English words: Some implications from Stroop task results. *Cognitive Psychology*, *11*, 125–132. http://dx.doi.org/10.1016/0010-0285(79)90007-0.

Birnbaum, I. M., & Eichner, J. T. (1971). Study versus test trials and long-term retention in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *10*, 516–521. http://dx.doi.org/10.1016/S0022-5371(71)80023-3.

van den Broek, G. S. E., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*, 803–812. http://dx.doi.org/10.1080/09658211.2013.831455.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*, 26–48. http://dx.doi.org/10.3758/BF03210724.

Dunn, J. C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review*, *115*, 426–446. http://dx.doi.org/10.1037/0033-295X.115.2.426.

Dunn, J. C., & James, R. N. (2003). Signed difference analysis: Theory and application. *Journal of Mathematical Psychology*, *47*, 389–416. http://dx.doi.org/10.1016/S0022-2496(03)00049-X.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*, 91–101. http://dx.doi.org/10.1037/0033-295X.95.1.91.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 840–859. http://dx.doi.org/10.1037/a0027867.

Heathcote, A., Freeman, E., Etherington, J., Tonkin, J., & Bora, B. (2009). A dissociation between similarity effects in episodic face recognition. *Psychonomic Bulletin & Review*, *16*, 824–831. http://dx.doi.org/10.3758/PBR.16.5.824.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 562–567. http://dx.doi.org/10.1016/S0022-5371(71)80029-4.

Hopper, W. J., & Huber, D. E. (2018). Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects. *Journal of Memory and Language*, *102*, 1–15. http://dx.doi.org/10.1016/j.jml.2018.04.005.

Jang, Y., & Nelson, T. O. (2005). How many dimensions underlie judgments of learning and recall? Evidence from state-trace methodology. *Journal of Experimental Psychology: General*, *134*, 308–326. http://dx.doi.org/10.1037/0096-3445.134.3.308.

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, *119*, 186–200. http://dx.doi.org/10.1037/a0025960.

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *Quarterly Journal of Experimental Psychology*, *65*, 962–975. http://dx.doi.org/10.1080/17470218.2011.638079.

Kalish, M. L., Dunn, J. C., Burdakov, O. P., & Sysoev, O. (2016). A statistical test of the equality of latent orders. *Journal of Mathematical Psychology*, *70*, 1–11. http://dx.doi.org/10.1016/j.jmp.2015.10.004.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology*, *93*, 329–343. http://dx.doi.org/10.2307/1422236.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370. http://dx.doi.org/10.1037/0096-3445.126.4.349.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, *131*, 147–162. http://dx.doi.org/10.1037/0096-3445.131.2.147.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*, 312–319. http://dx.doi.org/10.3758/BF03197461.

Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *11*, 397–406. http://dx.doi.org/10.1037/0278-7393.11.2.397.

Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review*, *111*, 835–865. http://dx.doi.org/10.1037/0033-295X.111.4.835.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *10*, 756–766. http://dx.doi.org/10.1037/0278-7393.10.4.756.

MacLeod, C. M., & Nelson, T. O. (1984). Response latency and response accuracy as measures of memory. *Acta Psychologica*, *57*, 215–235. http://dx.doi.org/10.1016/0001-6918(84)90032-5.

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*, 322–336. http://dx.doi.org/10.1037/0278-7393.31.2.322.

Millward, R. (1964). Latency in a modified paired associate learning experiment. *Journal of Verbal Learning and Verbal Behavior*, *3*, 309–316. http://dx.doi.org/10.1016/S0022-5371(64)80071-2.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109–133. http://dx.doi.org/10.1037/0033-2909.95.1.109.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*, 267–270. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00147.x.

Newell, B. R., Dunn, J. C., & Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition*, *38*, 563–581. http://dx.doi.org/10.3758/MC.38.5.563.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922. http://dx.doi.org/10.1162/neco.2008.12-06-420.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin*, *137*, 131–148. http://dx.doi.org/10.1037/a0021705.

Roediger, H. L. III., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210. http://dx.doi.org/10.1111/j.1745-6916.2006.00012.x.

Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition*, *22*, 511–524. http://dx.doi.org/10.3758/BF03198390.

Spellman, B. A., & Bjork, R. A. (1992). People's judgments of learning are extremely accurate at predicting subsequent recall when retrieval practice mediates both tasks. *Psychological Science*, *3*, 315–316. http://dx.doi.org/10.1111/j.1467-9280.1992.tb00680.x.

Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2018). Are there two processes in reasoning? The dimensionality of inductive and deductive inferences. *Psychological Review*, *125*, 218–244. http://dx.doi.org/10.1037/rev0000088.

Wearing, A. J., & Montague, W. E. (1970). A test of the Battig procedure for controlling the level of individual item learning in paired-associate lists. *Behavioral Research Methods and Instrumentation*, *2*, 9–10. http://dx.doi.org/10.3758/BF03205715.

Wescourt, K. T., & Atkinson, R. C. (1973). Scanning for information in long- and short-term memory. *Journal of Experimental Psychology*, *98*, 95–101. http://dx.doi.org/10.1037/h0034311.

Wilson, J. H., & Criss, A. H. (2017). The list strength effect in cued recall. *Journal of Memory and Language*, *95*, 78–88. http://dx.doi.org/10.1016/j.jml.2017.01.006.

Wixted, J. T., Ghadisha, H., & Vera, R. (1997). Recall latency following pure- and mixed-strength lists: A direct test of the relative strength model of free recall. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *23*, 523–538. http://dx.doi.org/10.1037/0278-7393.23.3.523.