# Disentangling Similarity Judgments from Pragmatic Judgments: Response to Sloutsky and Fisher

**Nicholaus S. Noles** and **Susan A. Gelman**
University of Michigan

## Abstract

Sloutsky and Fisher attempt to reframe the results presented in Noles and Gelman as a pure replication of their original work validating the Similarity, Induction, Naming, and Categorization (SINC) model. However, their critique fails to engage with the central findings reported in Noles and Gelman, and their re-analysis fails to examine the key comparison of theoretical interest. In addition to responding to the points raised in Sloutsky and Fisher's critique, we elaborate on the pragmatic factors and methodological flaws present in Sloutsky and Fisher (2004) that biased children's similarity judgments. Our careful replication of this study suggests that, rather than measuring the influence of labels on judgments of perceptual similarity, the original design measured sensitivity to the pragmatics of task demands. Together, the results reported in Noles and Gelman and the methodological problems highlighted here represent a serious challenge to the validity of the SINC model specifically and the words-as-features view more generally.

### Keywords

We are grateful for the opportunity to respond to the critique by Sloutsky and Fisher (henceforth, S&F) and to clarify the major contributions of Noles & Gelman (henceforth, N&G). In brief, we believe that S&F overlook N&G's primary results, which demonstrate powerful pragmatic influences on children's similarity judgments and call into question the validity of the Similarity, Induction, Naming, and Categorization (SINC) model. We also dispute the assumptions and approach of S&F's data-analytic critique. Finally, we provide a new model, based on pragmatic principles, that predicts the available data better than does the SINC model.

## What are the primary findings reported in N&G?

First, we think it is important to take a step back and focus on N&G's central findings, about which there is no disagreement. The primary findings in our report are the main effect of task order (discriminable items first [referred to as Order 1 in S&F] vs. identical items first [referred to as Order 2 in S&F]), and the interaction between order and similarity. S&F do not dispute these effects, and instead focus their critique on a 3-way interaction that is subsidiary to our main point. Our findings are of central interest because they support two important new conclusions: (a) Prior tests of the SINC model obscured powerful effects of order on children's similarity judgments, and (b) task demands in Sloutsky and Fisher's (2004) original work undermined children's ability to make perceptual similarity judgments, thereby greatly overestimating the effects of labels when they were available (see below).

## Do our findings undermine the labeling effect?

There are two different ways to ask the question of whether our findings undermine Sloutsky and Fisher's (2004) labeling effect. The first way is to ask: how much do children use labels in their similarity judgments, when labels are available? (Note that this version of the question focuses strictly on the label condition, and is agnostic regarding how the label and no-label conditions relate to one another.) Here the data clearly show that children use labels dramatically less often under the discriminable-first than the identical-first order. This is the finding to which we referred when we wrote: "For our purposes, the key finding is that children were much more accurate at making perceptual judgments in one task order than in the other" and "It is important to note that these data undermine the claim that labels contribute to the perceptual similarity of items (Sloutsky & Fisher, 2004), as labels influence perceptual similarity primarily under the adverse order of conditions." We also found that children in the discriminable-first order were more accurate in the no-label condition as well. This finding does not undercut the relevance of task order to judgments in the label conditions, as S&F repeatedly suggest, but rather underscores the powerful effects of task order in S&F's design. Providing identical item sets first undermined children's ability to make simple perceptual judgments, in all three conditions, thus yielding a significant main effect of order, and a significant interaction between order and similarity.

The second way of examining how our findings speak to the labeling effect is to ask: how much are children using labels, over and above the no-label condition? In our original paper, we reported no significant interaction with labeling condition ("…no significant three-way interaction involved condition…"), and we explicitly noted that our data replicated Sloutsky and Fisher's (2004) original labeling effect ("…our data replicate the results reported in Sloutsky and Fisher (2004), indicating that labels influence similarity judgments."). Thus, we have no quarrel with S&F's observation that the labeling factor did not interact with task order. However, it is nonetheless of clear theoretical interest to know if the previously reported label effect (higher similarity judgments in the label vs. no-label conditions) holds up in all cells of our design. In other words, we wished to examine the robustness of the label effect, particularly in the cell that provides the cleanest test of the hypothesis (i.e., the discriminable items in the discriminable-first order). To address this question, we conducted planned comparisons and discovered no statistically significant effects of label condition for the discriminable items in the discriminable-first order ($p$s = 1.0). In other words, when children are first presented with discriminable trials, they make accurate perceptual match judgments on the discriminable item sets whether provided with labels or not. S&F object to this analysis in part because they characterize it as an a posteriori test. However, it is explicit in both our criticism of their original work and in the design of the reported experiments that the discriminable items in the discriminable-first order provide the more appropriate test of the labeling effect, and thus we planned a priori to conduct this analysis. Conducting planned comparisons is a common and accepted practice in studies of developmental psychology, even in the context of non-significant interactions, as the authors themselves have done in prior published research (e.g., see Lawson & Fisher, 2011; Osth, Dennis, & Sloutsky, 2010; Sloutsky & Robinson, 2008 for recent examples).

## What is the theoretical status of the identical items?

S&F state that in triads where SR = 1, "both test items were equally similar to the target (N&G call these 'Identical' triads)." Importantly, however, it was not only that both test items were "equally similar to the target"; instead, in both Sloutsky and Fisher (2004) and N&G, all three pictures in a given triad were *exactly the same*. The items were identical. We maintain that *any* theory would predict that children would be likely to use labels for the identical items, because the labels were the only feature differentiating these pictures. This is

the task demand mentioned above, and conforms to basic pragmatic principles (Grice, 1975). For the identical sets, the fact that the only differentiating feature is the label encourages the inference that the label is relevant to the task at hand. Accordingly, children are biased to make use of the label, in order to be a cooperative interlocutor. This problem arises not only with the identical sets, but also, we contend, for most of the items in Sloutsky and Fisher's (2004) Experiment 1, as children's performance in the no-label condition was at or near chance for SRs = 1, 1.22, and 1.86 (three of the four SRs).

Moreover, even in the no-label condition, the identical sets provide an impossible task, thereby encouraging a pure-guessing strategy (not just on the identical trials but also carrying over to later trials where in fact the items are discriminable). Thus, the only appropriate test of the SINC theory is to examine the *discriminable* items when they are presented first (uncontaminated by the biased identical sets). Only this cell of the design tests the model, because it is the only cell that is not susceptible to the alternative explanatory account laid out in N&G. S&F's reanalysis finding a significant labeling effect (reported in their commentary as an effect in the "non-adverse Order 1") collapses over identical and discriminable item sets, and thus does not test whether labeling influences children's judgments in the absence of pragmatic biases (namely, the discriminable sets only). When one specifically conducts a post-hoc comparison on the *discriminable* item sets in the discriminable-first order, collapsed over the two different labeling conditions (as S&F had done), there is still no significant difference in similarity judgments in the label vs. no-label conditions ($p > .30$).

A further potential problem with both types of items (identical and discriminable) is that in Sloutsky and Fisher's (2004) original work (though not N&G's replication), it appears that some base images were used in multiple item sets (e.g., the same dog and lion might be used to manufacture item sets with similarity ratios of both 9 and 1.22), creating the possibility that children might receive an item as a lure in one trial and a target in a subsequent trial, or vice versa. The possibility exists that stimulus repetition influenced perceptual judgments in Sloutsky and Fisher's original study. More generally, as can be seen in N&G Figure 1, the base images used to create Sloutsky and Fisher's stimuli did not always yield triads that were judged by adults to be equally similar, despite ostensibly sharing the same SR. The implications of these non-standard design decisions remain untested.

## Is critical information missing from N&G?

S&F criticize N&G for failing to report the effect size ($\eta_p^2$) of a non-significant difference in label use for one presentation order. Importantly, the effect size that S&F singled out collapses over identical and discriminable item sets and therefore is not relevant to the key comparison of interest: performance on discriminable trials when they are presented first. In any case, as noted above, the effect size in question was non-significant. Because even large effect sizes do not necessarily translate to significant effects, we limited our comments to a discussion of the tests of statistical significance, as reflected in the $p$-values.

S&F also present a similar argument in the form of a table reporting Cohen's d for a subset of our data. These values reflect moderate to large effects, but the overall character of these data is obfuscated by the selective presentation of information. As can be seen in Table 1, two important points become apparent when all of the pertinent information is presented. First, most of the comparisons discussed by S&F are non-significant. Second, use of the label increases in the identical-first order as compared to the discriminable-first order, in some cells rather dramatically.

## How well does the SINC model predict the data?

In their commentary, S&F state: "N&G data perfectly replicate S&F findings even in the most critical Order 1." This statement is based on *predicting* our results, using an estimate of w derived from a *subset* of our data (Label Plus Training, Order 1). Instead of underlining predicting our results based on a underline subset of our data, we believe it is more appropriate to compare their predictions to our underline actual findings, from underline all eight cells of our design (2 labeling conditions × 2 orders × 2 similarity ratios). For the reasons set forth in the section entitled "What is the theoretical status of the identical items?", the most relevant comparisons concern the discriminable items (SR=9), although discrepancies between S&F and our data also appear when examining lower SR values.[i]

As can be seen in Figure 1, our data seriously deviate from the predictions made by the SINC model. To orient the reader: w=.35 is S&F's prediction for the label conditions, and w=1.0 is their prediction if there is no effect of labeling. The remaining bars present our data from the labeling conditions (LPT, LMT) and the no-label condition (NL), under two different presentation orders. First, we find a large order effect that is not predicted by the SINC model (compare discriminable-first order vs. identical-first order). Second, in the discriminable-first order, scores in the label conditions do not significantly differ from scores in the no-label condition. It is only in the identical-first (contaminated) order that labeling has a significant effect. This outcome is inconsistent with the basic assumptions of the SINC model. Third, in the discriminable-first order, scores in the label conditions do not significantly differ from the SINC prediction of no-labeling effects (i.e., w=1), where the estimate is .10 ($ps$ = .49 and .12 in the LPT and LMT conditions, respectively; $p$ = .17 when the conditions are collapsed). For all these reasons, we conclude that N&G's data do not support S&F's predictions.

In contrast, a pragmatic model does a much better job of predicting the data. We constructed a rather crude pragmatic model—certainly over-simplified and not sufficient for all future cases, but one that follows from the following simple assumptions about the source and nature of children's similarity judgments:

- We assume that children's responses include a combination of two types of trials: those on which children correctly perceive the correct choice, and those on which children are guessing.

- Guessing trials in the no-label conditions are estimated to be 2*P(B), given chance responding of 50% (e.g., if a child is 90% correct, then P(B) is .10, and guessing is .20). P(B) rates in the no-label condition can be straightforwardly derived from the similarity ratios in S&F, and are presented in Table 2.

- **r** is defined as the rate at which children default to using a label on **guessing trials** (estimated from the data in both S&F and N&G to be .75). This is assumed to be a constant rate.

- **G** (for pragmatic, or Gricean bias) is the rate at which children default to using a label on **non-guessing trials**. This will be multiplied by the rate at which the context presents a pragmatic bias (s), for example due to an identical-first order. Note that G will vary across different experimental contexts. Contexts with a strong pragmatic bias will have a higher G than contexts with a weak pragmatic bias. In

---

[i]S&F's estimates of w in their Table 2 are incomplete in two respects. First, they are based on only 2 of the 8 cells in N&G's study design. When all 8 cells are included, estimates of w vary more widely, ranging from .21 to .50. Second, w can alternatively be calculated by entering our obtained values of P(B) (from the naming conditions) into Equation 5 of Sloutsky and Fisher (2004), yielding estimates of w that range from .08 to .58. Thus, the predictions presented in S&F's Figure 1 are artificially restricted.

N&G, G can be estimated by comparing biased (identical-first) to unbiased (discriminable-first) presentation orders for non-guessing (discriminable) trials. In S&F, G can be estimated from best fit to the data.

- **s** is defined as the presence or absence of pragmatic bias. If all discriminable items appear first, s=0; otherwise, s=1. In N&S, s=0 in Order 1; s=1 in Order 2. For S&F's original experiment, s is also estimated to be approximately 1, because discriminable and non-discriminable sets were intermixed over 16 trials. (Note that the majority of S&F's trials were non-discriminable, even for SR=1.22 and SR=1.86.)

- Based on these assumptions, the simple pragmatic model is:

$$[2^{*}P(B)_{no-label}{}^{*}r] + [(1 - (2^{*}P(B)_{no-label}))^{*}G^{*}s]$$

Table 2 presents the predictions from SINC and the Pragmatic model, and Figures 2(a) and 2(b) compare these predictions to the obtained data. As can be seen, even a very simple Pragmatic model better predicts the data than SINC. (We assume that a more realistic pragmatic model would be more complex, given the subtlety involved in pragmatics.)

## Conclusions

It seems that four firm conclusions follow from N&G's experiments and the current exchange: (1) the SINC model predicts children's similarity judgments no better than a model based on rates of guessing adjusted for pragmatic bias, (2) S&F's primary defense of the SINC model relies on discussing effects that are non-significant, (3) there is no evidence of an influence of shared labels on children's similarity judgments in the "critical cell" of N&G's design (discriminable items in the discriminable-first condition), and (4) S&F do not present a compelling argument against, or do more than briefly mention, the central findings presented in N&G (i.e., the significant effects of order and the interaction of order and similarity). Given these problems, how can S&F conclude that our data support their similarity-based account?

We argue that Sloutsky and Fisher's (2004) original test was methodologically flawed in terms of both item selection and pragmatic biases. N&G provide new evidence that was obscured by Sloutsky and Fisher's (2004) original design and analyses. There is now substantial evidence in the literature that pragmatic factors can direct young children's performance on cognitive and language tasks (e.g., Diesendruck & Markson, 2001; Grassmann, Stracke, & Tomasello, 2009), and our data are consistent with this view. Returning to the broader question of what N&G's findings tell us about children's concepts, we find little support for the words-as-features view or the position that labels exercise influence over the perceived perceptual similarity of entities above and beyond the task demands inherent in Sloutsky and Fisher's experimental methods. We therefore suggest that other explanations are required for understanding the relations among language, concepts, and induction.

## Acknowledgments

# References

Diesendruck G, Markson L. Children's avoidance of lexical overlap: A pragmatic account. Developmental Psychology. 2001; 37:630–641.10.1037/0012-1649.37.5.630 [PubMed: 11552759]

Grassmann S, Stracke M, Tomasello M. Two-year-olds exclude novel objects as potential referents of novel words based on pragmatics. Cognition. 2009; 112:488–493.10.1016/j.cognition.2009.06.010 [PubMed: 19616205]

Grice, HP. Logic and conversation. In: Cole, P.; Morgan, J., editors. Syntax and semantics 3: Speech arts. New York: Academic Press; 1975. p. 41-58.

Lawson CA, Fisher AV. It's in the sample: The effect of sample size on the breadth of inductive generalization. Journal of Experimental Child Psychology. 2011; 110:499–519.10.1016/j.jecp. 2011.07.001 [PubMed: 21816410]

Noles NS, Gelman SA. Effects of categorical labels on similarity judgments: A critical analysis of similarity-based approaches. Developmental Psychology. 2012; 48:xxx–xxx. doi:10.1037/ a0026075.

Osth, AF.; Dennis, S.; Sloutsky, VM. Context and category information in children and adults. In: Ohlsson, S.; Catrambone, R., editors. Proceedings of the XXXII Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum; 2010. p. 842-847.

Sloutsky VM, Fisher AV. Induction and categorization in young children: A similarity-based model. Journal of Experimental Psychology: General. 2004; 133:166–188.10.1037/0096-3445.133.2.166 [PubMed: 15149249]

Sloutsky VM, Fisher AV. Effects of categorical labels on similarity judgments: A critical evaluation of a critical analysis (Comment on Noles and Gelman, 2012). Developmental Psychology. 2012; 48:xxx–xxx. doi:10.1037/a0027531.

Sloutsky VM, Robinson CW. The role of words and sounds in infants' visual processing: From overshadowing to attentional tuning. Cognitive Science. 2008; 32:354–377.10.1080/03640210701863495
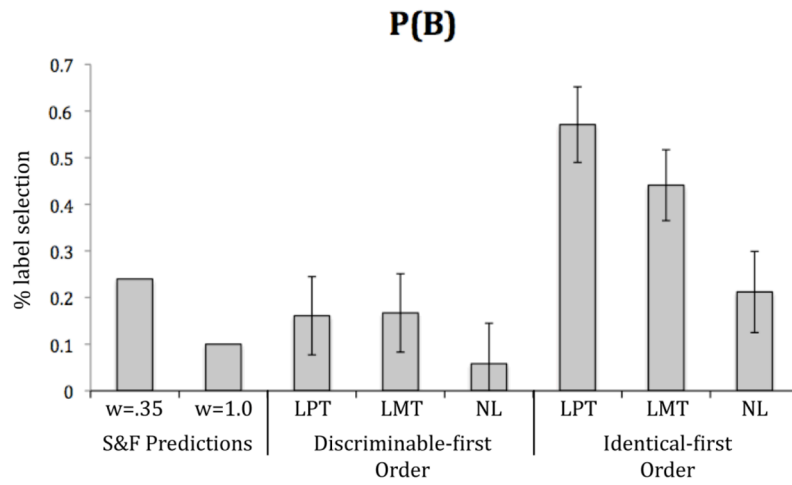
**Figure 1.**
P(B) (mean % label selections) for discriminable items only (SR = 9), as predicted by
Sloutsky & Fisher (2004) (label effects, w=.35; no effect of label, w=1.0) and as obtained by
Noles & Gelman (2012) (Discriminable-first Order and Identical-first Order). Note: LPT =
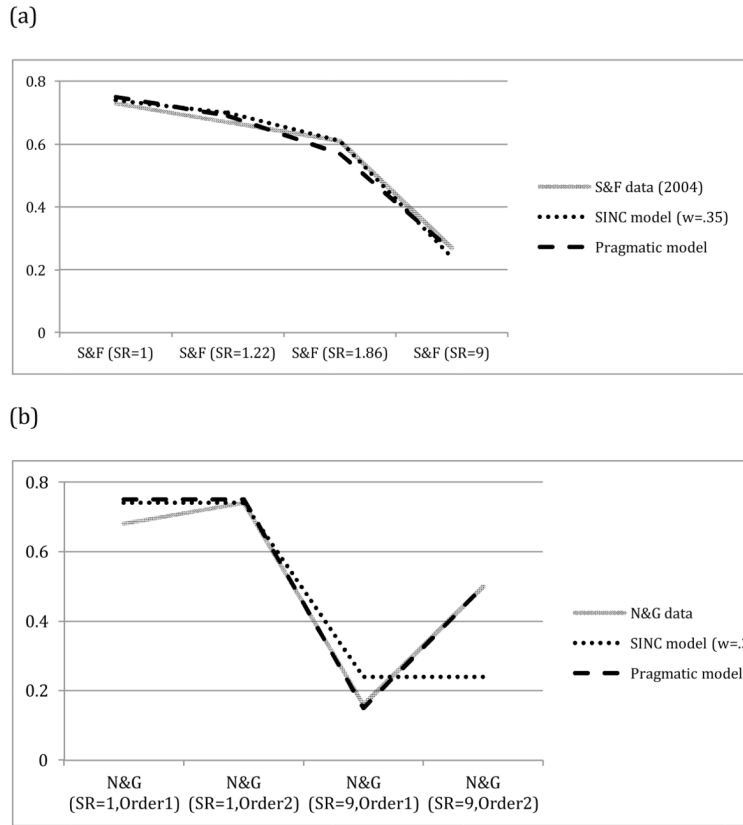Labels plus training; LMT = Labels minus training; NL = No labels.

(a)



(b)



**Figure 2.**
(a) Predictions from the SINC model and from the Pragmatic model, as compared to the results from Sloutsky & Fisher (2004), Experiment 1. Note: Data from S&F were calculated from Figure 7A in Sloutsky & Fisher (2004). There are small discrepancies between the values presented in Figure 1 of S&F's current commentary and those presented in Figure 7A of Sloutsky & Fisher (2004). (b) Predictions from the SINC model and from the Pragmatic model, as compared to the results from Noles & Gelman (2012). Note: "Order 1" refers to the discriminable-first trials; "Order 2" refers to the identical-first trials.

**Table 1**

Mean differences, SEs, $p$-values, and Cohen's d for label vs. no-label comparisons, Noles & Gelman (2012).

| | Mean difference \|label – nolabel\| | SE | $p$-value | Cohen's d |
|---|---|---|---|---|
| DISCRIMINABLE ITEMS[a] | | | | |
| Discriminable-first order[b] | | | | |
| LPT vs. NL | .10 | .12 | 1.00 | 0.43 |
| LMT vs. NL | .11 | .12 | 1.00 | 0.81 |
| Identical-first order[c] | | | | |
| LPT vs. NL | .36 | .12 | .01* | 0.93 |
| LMT vs. NL | .23 | .12 | .15 | 0.64 |
| IDENTICAL ITEMS[d] | | | | |
| Discriminable-first order[b] | | | | |
| LPT vs. NL | .20 | .11 | .20 | 0.80 |
| LMT vs. NL | .17 | .10 | .34 | 0.62 |
| Identical-first order[c] | | | | |
| LPT vs. NL | .32 | .11 | .01* | 1.24 |
| LMT vs. NL | .18 | .10 | .25 | 0.63 |

Note 1: LPT = Labels plus training; NL = no labels; LMT = Labels minus training.

Note 2: We calculated Cohen's d based on means and SDs; some of our estimates differ from those provided by S&F.

[a] corresponds to "SR = 9" in S&F

[b] corresponds to "Order 1" in S&F

[c] corresponds to "Order 2" in S&F

[d] corresponds to "SR = 1" in S&F

**Table 2**

Predictions made by the SINC model and the Pragmatic model, for the label conditions in Sloutsky & Fisher (2004) and Noles & Gelman (2012).

| Data Source | SR | P(B) no label | P(guessing) | r | G | s | SINC prediction | Pragmatic prediction |
|---|---|---|---|---|---|---|---|---|
| S&F (2004) | 1 | 0.50 | 1.00 | 0.75 | 0.14 | 1 | 0.74 | 0.75 |
| S&F (2004) | 1.22 | 0.45 | 0.90 | 0.75 | 0.14 | 1 | 0.70 | 0.69 |
| S&F (2004) | 1.86 | 0.35 | 0.70 | 0.75 | 0.14 | 1 | 0.61 | 0.57 |
| S&F (2004) | 9 | 0.10 | 0.20 | 0.75 | 0.14 | 1 | 0.24 | 0.26 |
| N&G-Order1 | 1 | 0.50 | 1.00 | 0.75 | 0.44 | 0 | 0.74 | 0.75 |
| N&G-Order2 | 1 | 0.50 | 1.00 | 0.75 | 0.44 | 1 | 0.74 | 0.75 |
| N&G-Order1 | 9 | 0.10 | 0.20 | 0.75 | 0.44 | 0 | 0.24 | 0.15 |
| N&G-Order2 | 9 | 0.10 | 0.20 | 0.75 | 0.44 | 1 | 0.24 | 0.50 |

Note: "Order 1" refers to the discriminable-first trials; "Order 2" refers to the identical-first trials.