Combination or Differentiation? Supplementary Materials

Andy J. Wills[a,*], Angus B. Inkster[a], Fraser Milton[b]

[a]*School of Psychology, Plymouth University, UK*
[b]*Psychology, University of Exeter, UK*

Abstract

This document contains supplementary materials for the main paper. Section 1 reports additional response-set analyses of the experiments in the main paper (with further detail in the Appendix). Section 2 reports an alternative, logistic-regression-based, analysis of the criterial-attribute experiments. Section 3 reports analyses of the impulsivity self-report data collected in Experiments 2, 3A and 3B. Section 4 considers whether the conclusions of the main paper would have been different if we had chosen to replicate different studies from the same classic papers (Kemler Nelson, 1984; Smith & Kemler Nelson, 1984; Smith & Shapiro, 1989; Ward, 1983).

1. Consistency and margin

In the main paper, our analysis of Experiments 1–3B focussed on the proportion of participants classified as Overall Similarity (OS), Identity (ID), and Unidimensional (UD) responders. Similarly, our analysis of Experiments 4–5 focussed on the proportion of participants classified as Overall Similarity (OS), Criterial Attribute (CA), or Non-criterial Attribute (NCA) responders. In this Section, we consider two further dependent variables (in addition to proportion) that response-set analysis can provide — *consistency* and *margin*. Recall that response-set analysis calculates, for each participant, the proportion of responses predicted by each response strategy. The strategy that predicts the highest proportion of responses is selected. *Consistency* is the proportion of responses predicted by the selected strategy. *Margin* is difference between the consistency of the selected strategy and the consistency of the strategy in second place. For example, if the proportion of responses predicted for a particular participant were .8 for the UD strategy, .4 for the OS strategy, and .6 for the ID strategy, then *consistency* for that participant would be .8 and the *margin* for that participant would be $.8 - .6 = .2$.

Consistency and margin can be used in both a descriptive and an inferential manner, and both approaches are employed in the sections that follow. Descriptively, one can ask how good a selected response strategy is in absolute terms, and *consistency* gives us a measure for this. If the best available strategy captures a low proportion of the participants' responses,

*Corresponding author
Email address: andy@willslab.co.uk (Andy J. Wills)

this might prompt one to consider whether there are other response strategies, not included in the analysis, which might have done better. One can also consider the extent to which the selected strategy does a better job of predicting the participant's responses than the other included but non-selected strategies. If *margin* is small, it would be inappropriate to attach much importance to the absolute proportion of participants classified as using each strategy, although differences in proportion between experimental conditions should still be interpretable.

Inferentially, one can examine whether one response strategy, over all the cases it is selected, captures a higher proportion of responses than another strategy (*consistency*) or wins over its competitors more than another strategy does (*margin*). One can also examine whether an experimental manipulation (e.g. time pressure) affects how consistently the selected strategy was employed (*consistency*), and whether that manipulation affects the extent to which the strategies are distinguishable (*margin*).

Consistency and margin were both substantial in the triad task. Across all participants in Experiments 1, 2, 3A and 3B, excluding the position-bias response models, the best-fitting response model predicted a mean of 74% of responses (against a chance level of 33%). The best-fitting model predicted, on average, 23% more responses than the next best model (i.e. the second-best model predicted, on average 51% of the participant's responses).

Measure	UD	OS	ID
consist.	.71	.70	.84
margin	.19	.21	.31
N	147	139	98

Table 1: Further response-set analyses for the triad-task experiments. Consist.: mean proportion of responses predicted by the best-fitting model. Margin: mean difference in proportion of responses predicted by the best-fitting and second-best-fitting model. N: sample size

Inspection of Table 1 suggests that the ID model had higher consistency than the OS and UD model, which did not differ. Statistical analysis confirms this impression, $F(2, 381) = 36.09, p < .0001$, with a Tukey HSD test reporting $p = .98$ for the UD-OS comparison, and $p < .0001$ for the other two pairwise comparisons. The same pattern is seen for margin, $F(2, 381) = 26.95, p < .0001$, with Tukey HSD $p = .22$ for the UD-OS comparison and $p < .0001$ for the other two pairwise comparisons. In terms of contrasting the predictions of Differentiation and Combination Theory, it is the relative prevalence of OS and UD responders that is of most importance. It is thus reassuring that these two models are comparable both in the proportion of responses they predict when they win, and the extent to which they beat their nearest competitor when they win. This contrasts with, for example, General Recognition Theory analysis of information-integration category learning experiments (Edmunds et al., accepted).

Turning to the criterial-attribute task, consistency and margin were both substantial. Across all participants in Experiments 4A, 4B, and 5, excluding the position-bias response models, and removing OS-ambiguous trials from the analysis (see main paper), the best-

fitting response model predicted a mean of 86% of responses (against a chance level of 50%). The best-fitting model predicted, on average, 13% more responses than the next best model (i.e. the second-best model predicted, on average, 73% of the participant's responses).

Measure	OS	CA	NCA
consist.	.77	.92	.80
margin	.03	.16	.12
N	19	106	65

Table 2: Further response-set analyses for the criterial-attribute experiments. Consist.: mean proportion of responses predicted by the best-fitting model. Margin: mean difference in proportion of responses predicted by the best-fitting and second-best-fitting model. N: sample size

Table 2 indicates that the CA model had higher consistency than the OS and NCA model, which did not differ. Statistical analysis confirms this impression, $F(2, 187) = 25.94, p < .0001$, with Tukey HSD $p = .56$ for the OS-NCA comparison, and $p < .0001$ for the other two pairwise comparisons. In terms of comparing the predictions of Combination and Differentiation Theory, the critical question concerns whether, when people switch away from criterial-attribute responding, they switch to OS responding, or NCA responding. It is thus reassuring that the OS and NCA models are comparable in the proportion of responses they predict when they win.

Margin is much lower for the OS model (.03) than the NCA model (.12). Again, statistical analysis confirms this conclusion, $F(2, 187) = 48.96, p < .0001$, with Tukey HSD $p < .0001$ for all three pairwise comparisons. Thus, even when the OS model wins (a rare event), the extent to which it is a better predictor of participant's behaviour than its closest competitor is very small. One should probably not read much into the fact that a few participants were classified as OS responders, as some of the already-small number of OS "wins" are likely to be driven by noise rather than signal. In summary, these analysis indicate that the concept of Overall Similarity responding has low explanatory value in the context of the criterial-attribute task.

It is also the case that the CA model wins by a larger margin than the NCA model. This is likely to be, at least in part, a side-effect of the increased proportion of NCA responders in conditions where cognitive resources are limited (incidental training, concurrent load). Consistency is lower in such conditions, which in turn tends to reduce the margin of win. However, in Experiment 5, the CA model had a higher margin than the NCA model, even under conditions where there was no detectable effect of concurrent load on consistency or margin (those conditions being when only passing participants were analysed and OS-ambiguous stimuli were included). It is not possible to tell from these analyses whether this difference is due to the NCA model being a less good model of human behaviour than the CA model, or whether those applying an NCA strategy simply have more difficulty doing so consistently than those applying a CA strategy.

The analyses presented in this section are available at `www.willslab.co.uk/plym9/`. In the Appendix, we analyze each experiment separately, and report consistency and margin

for each combination of model and experimental condition.

2. Logistic regression analysis of the criterial-attribute experiments

In the main paper, we considered response models where one stimulus dimension had exclusive control over responding (the CA and NCA models), and a response model where all four stimulus dimensions had equal control over responding (the OS model). There are a number of other possible response strategies that are not captured in this relatively simple analysis. For example, a participant may use more than one, but less than four, stimulus dimensions in their decision. Although this could be addressed by adding 2- and 3-dimension models to the main analysis, such an approach would not address the possibility that participants weight the stimulus dimensions unequally (e.g. giving more weight to the moustache than the ears). Below, we address both these limitations by using an analysis method based on logistic regression.

A useful feature of the test phases of Experiments 4A and 5 is that all 16 stimuli it is possible to create with four binary-featured stimuli are presented with equal frequency. This means that, in the context of the test phase, the stimulus features are uncorrelated with each other. In turn, this opens the possibility of using the co-efficients of a multiple regression analysis to assess, for each participant, the extent to which each stimulus dimension predicts their responses. We have previously used this procedure in related studies of overall similarity classification (Lea et al., 2009). In the current case, we employed logistic regression because the dependent variable (categorization response) was binary. More specifically, we employed Firth's method of penalized-likelihood logistic regression (Firth, 1993), using the `logistf` package (Henize et al., 2013) in the R software environment (R Core Team, 2014). Firth's method addresses a problem that occurs in standard logistic regression if there are perfect or near-perfect predictors of the dependent variable. This was the case in the current analysis—for some participants, a single stimulus dimension perfectly predicted their responses.

Using Firth's method, we derived four regression coefficients for each participant (one coefficient for each stimulus dimension). This set of regression coefficients is a rich data set, amenable to a range of analyses. In order to remain close to the hypotheses and DVs of the main paper, we used the regression coefficients to derive two simple dependent variables. The *dimensions* DV indexed the number of stimulus dimensions that had a detectable effect on participant's responses, defined as the number of regression coefficients significant at .05, uncorrected, for that participant. The *criterial* DV was binary, indicating for each participant whether the stimulus dimension with the greatest effect on responding (defined as the dimension with largest absolute regression coefficient) was the criterial dimension, or a non-criterial dimension.

Participants with zero significant coefficients (i.e. participants whose responses could not be significantly predicted from any linear combination of the stimulus features) were excluded from all analyses. In Experiment 4A, ten participants were excluded for this reason. Inspection of Table 3 indicates that single-dimension responding was the modal strategy in both conditions, supporting the conclusion of the main analysis that single-dimension responding was prevalent under both intentional and incidental training conditions. Very

		Dimensions				Criterial
		1	2	3	4	
Exp. 4A	Intentional	.66	.19	.09	.06	.72
	Incidental	.50	.30	.16	.04	.28
Exp. 5	Concurrent load	.87	.08	.05	.00	.34
	Full attention	.81	.12	.05	.02	.71

Table 3: Proportion of participants with 1–4 dimensions significantly affecting their response("dimensions"), and proportion of participants for whom the criterial attribute is the best predictor of their response ("criterial"), reported for each condition of Experiments 4A and 5.

few participants used all four dimensions, confirming the conclusion that overall-similarity classification is rare in this task.

One notable aspect of the results in Table 3 is that a substantial minority of participants in both conditions used more than one but less than four dimensions in their classification. These participants' responses are not fully captured by any of the response models employed in the main paper. Nevertheless, the current analysis leads to the same conclusions as before. In the main paper, we reported that participants in the intentional condition were more likely to classify on the basis of the criterial-attribute than participants in the incidental condition. Correspondingly, the criterial-attribute is the best predictor of responding for 72% of participants in the intentional condition, but only 28% of participants in the incidental condition. This difference is significant, $\chi^2(1) = 16.68, p < .0001$. In the main paper, we further reported that incidental training did not produce any shift away from single-dimensional responding towards overall-similarity responding. Correspondingly, experimental condition did not significantly affect the distribution shown in Table 3, $\chi^2(3) = 2.64, p = .49$. Bayesian contingency-table analysis provides substantial evidence in support of the null, $BF = .18$.

Logistic regression cannot be validly applied to the data from Experiment 4B because the removal of OS-ambiguous stimuli from the test phase means that the stimulus dimensions are now correlated with each other, violating an assumption of the analysis. Logistic regression analysis can be applied to Experiment 5, and the conclusions are similar to those for Experiment 4A. Single-dimension responding was the modal strategy in both conditions, and very few participants used four dimensions (see Table 3). In both conditions, about 15% of participants used more than one, but less than four, dimensions. This is a smaller proportion than in Experiment 4A, meaning that the response-set models used in the main paper were a better approximation to the response strategies in Experiment 5 than in Experiment 4A. As might be expected under such conditions, the logistic regression analysis of Experiment 5 led to the same conclusions as the analysis in the main paper. Specifically, the proportion of participants for whom the criterial-attribute dimension was the best predictor of their responding was higher in the full-attention condition than the concurrent-load condition, $\chi^2(1) = 10.56, p = .001$, and experimental condition had no detectable effect on the distribution shown in Table 3, $\chi^2(3) = 1.39, p = .89$. Bayesian contingency-table analysis provided substantial evidence in support of the null, $BF = .22$.

The analyses reported in this section are available in the same data archives as the main analyses of the corresponding experiments (see main paper for details).

3. Impulsivity analysis

Ward (1983, Experiment 1) reported that impulsive participants produced a higher proportion of "overall similarity" (BC) responses than reflective participants. This result is consistent with Differentiation Theory under the assumption that impulsive participants are less willing or able to expend cognitive resources on the classification task than reflective individuals, and that BC responding reflects an overall similarity rather than a unidimensional strategy. It is inconsistent with Combination Theory under the same assumptions. A secondary purpose of Experiments 2, 3A and 3B was to re-examine the relationship between impulsivity and classification behavior in the triad task.

Ward (1983) employed the Matching Familiar Figures measure of impulsivity (Kagan, 1965), a measure whose validity has been questioned (e.g. Block, Block & Harrington, 1974) and which appears to be largely uncorrelated with better validated measures of impulsivity (Helmers, Young & Pihl, 1995). In the current study, we employed the Barratt Impulsivity Scale (BIS-11), which is the most widely used measure of impulsivity (Stanford, Mathias, Dougherty, Lake, Anderson & Patton, 2009). It has high reliability and good external validity (Patton, Stanford & Barratt, 1995). The Barratt Impulsivity Scale is a self-report measure that includes statements such as, "I concentrate easily" and "I am happy-go-lucky". Like Ward, our assessment of impulsivity immediately followed the triad classification task.

Ward's original demonstration employed an unspeeded procedure, so we focussed initially on the low time pressure conditions of Experiments 2 and 3A (impulsivity data were not collected in Experiment 1, or in the low time pressure condition of Experiment 3B). Traditional analysis revealed no significant correlation between impulsivity and BC responding in either Experiment 2, $r(39) = -.19, p = .24$, or Experiment 3A, $r(50) = -.20, p = .15$. Response-set analysis of Experiment 3A also revealed no significant effects, max. $t(39) = 1.55, p = .13$. Response-set analysis could not be performed in the low time pressure condition of Experiment 2 due the dominance of Identity responders in this sample.

Turning to the high time pressure conditions, no significant correlations between impulsivity and BC responding were revealed in Experiment 2, $r(37) = .08, p = .64$, Experiment 3A, $r(53) = .00, p = .98$, or Experiment 3B, $r(38) = -.27, p = .09$. If one wished to consider the last of these results as significant, despite the .1 alpha level, then it should be noted that the correlation (like all but one of the others) is negative, and thus implies BC responding reduces as impulsivity increases. This is opposite to Ward's result, and seems more consistent with Combination Theory than Differentiation Theory. Response-set analysis of the high time pressure condition of Experiment 3B is not possible due to the dominance of Overall Similarity responders in this sample. For similar reasons, only Unidimensional and Overall Similarity responders can be compared in Experiment 3A (they do not differ significantly, $t(50) < 1$) and only Unidimensional and Identity responders can be compared in Experiment 2 (they do not differ significantly either, $t(30) = 1.18, p = .25$).

We considered the possibility of combining across Experiments and conditions. There are five conditions for which we have impulsivity data: (1) Exp. 2, low time pressure (LTP), (2) Exp. 2, high time pressure (HTP), (3) Exp. 3A, LTP, (4) Exp. 3A, HTP, and (5) Exp. 3B, HTP. In our procedure (and in Ward's), the assessment of impulsivity follows the classification phase. Therefore, there is the possibility that the different classification procedures in these five conditions may have differentially affected impulsivity ratings. This kind of effect would not be entirely unprecedented (c.f. Verbruggen, Adams & Chambers, 2012), and there was some evidence that it had occurred in our studies. Across the ten possible pairwise comparisons of our five conditions, there were two significant differences in average impulsivity, both of which survived a Bonferroni correction for multiple comparisons. Specifically, average impulsivity was lower in Exp. 3A LTP (64.2) than in Exp. 2 LTP (71.6), $t(91) = 3.13, p = .002$, uncorrected, or in Exp. 2 HTP (70.7), $t(89) = 2.99, p = .004$, uncorrected.

Given these differences, any combination across conditions that included Exp. 3A LTP would introduce a confound that would render any observed effects largely uninterpretable. To illustrate the nature of this confound, observe that in Exp. 2 LTP, 90% of participants are Identity responders while in Exp. 3A LTP, only 18% of participants are Identity responders. In an analysis combining these two conditions, most Identity responders would come from Exp. 2A LTP, which has higher average impulsivity than Exp. 3A LTP. Symmetrically, most Overall Similarity responders in a combined analysis would come from Experiment 3A, which has lower average impulsivity. Differences in average impulsivity are thus confounded with the procedural differences between these Experiments, making any results of a combined analysis largely uninterpretable.

In our combined analysis, we thus set aside the data from Exp. 3A LTP and combined across the remaining four conditions. Average impulsivity did not significantly differ between these conditions. Traditional analysis revealed a significant negative correlation between impulsivity and BC responding in this combined data set, $r(173) = -.18, p = .02$. In other words, BC responding fell as impulsivity increased, a result that seems more consistent with Combination Theory than Differentiation Theory. Response-set analyses of the combined data set revealed no significant effects. The effect closest to significance was higher mean impulsivity for Identity responders (71.1) than for Overall Similarity responders (68.1), $t(125) = 1.62, p = .11$. This direction of effect is inconsistent with both Differentiation Theory and Combination Theory, but as the difference is not significant, it would be unwise to draw any strong conclusions on this basis.

In summary, using a well-validated measure of impulsivity (BIS-11), we found no evidence that overall similarity responding was related to high impulsivity, and traditional analyses provided some support for the opposite conclusion. Response-set analyses, while somewhat equivocal, provide no support for the contention that Overall Similarity responders are more impulsive than Identity or Unidimensional responders. It is possible that Ward's original results concerning impulsivity and overall similarity were an artifact either of the low validity of the measure of impulsivity employed, or a Type II error resulting from the use of a median split method of analysis (see MacCallum et al., 2002, for a critique of this method). However,

in the absence of further research, such a conclusion remains tentative. Future research on this topic would ideally collect both pre- and post-classification measures of impulsivity, rather than just the post-classification measures collected by Ward and ourselves. For the present, there seems to be little evidence from impulsivity measures in our experiments that would favor Differentiation Theory over Combination Theory.

Raw data for these analyses are available as part of the raw data archives for Experiments 2, 3A and 3B (see main paper). The analysis script for the impulsivity analyses is available at `www.willslab.co.uk/plym4/`.

4. Other experiments in the classic papers

In the current series of experiments, we focussed on large-scale replications of four experiments from four classic papers: Experiment 2 of Ward (1983), Experiment 3 of Smith and Kemler Nelson(1984), Experiment 1 of Kemler Nelson (1984) and Experiment 1 of Smith and Shapiro (1989). Below, we consider whether the outcome of our investigations would have been different if we had selected different experiments from within these multi-experiment papers.

4.1. Ward (1983)

Ward (1983) comprises three experiments. Experiment 2, the experiment we chose to replicate, is the only experiment in the paper to manipulate time pressure. Experiment 1 was a correlational study, looking at the relationship between classification style and total sort time in a self-paced task. Experiment 2's use of the experimental method makes it a better choice for replication. Experiment 3 was a developmental study that did not manipulate time pressure. Although developmental data is somewhat outside the scope of the current article, we note that response-set analyses of developmental triad task data have cast doubt on the conclusion that younger children are more likely to be overall similarity responders than older children or adults (Thompson, 1994; Raijmakers, Jansen & van der Maas, 2004).

4.2. Smith and Kemler Nelson (1984)

Smith and Kemler Nelson (1984) report six experiments. The results of Experiment 1, as the authors accept, are largely consistent with Combination Theory. In Experiment 1, participants are taught to classify triads in a particular way using trial-by-trial feedback. In the Similarity condition, feedback is consistent with an overall similarity strategy. In the Dimensional condition, feedback is consistent with a unidimensional strategy. Participants in the Similarity condition get both size-identity and brightness-identity triads, while participants in the Dimensional condition get only size-identity (or only brightness-identity) triads. It is not possible to train an identical-attribute strategy in the Dimensional condition using these stimuli; thus our best guess is that participants in the Dimensional condition are largely classifying on the basis of a unidimensional strategy. Participants take about twice as long to reach an errorless criterion in the Similarity condition than in the Dimensional condition. Participants also make significantly more errors on a subsequent low time pressure

test in the Similarity condition than the Dimensional condition. Both results seem to indicate that overall similarity classification is more difficult than unidimensional classification, which is consistent with Combination Theory. At the highest levels of time pressure, there is a nonsignificant trend for more errors in the Dimensional condition than the Similarity condition, which does seem more consistent with Differentiation Theory than Combination Theory—and this is the result on which the authors focus. However, the nonsignificance of the effect, the fact that the two conditions used different stimulus sets, and the fact that time pressure was manipulated within-subjects with higher time pressure always following lower time pressure (thus confounding time pressure with practice), lead us to the conclusion that this effect awaits replication, and that any attempt to replicate should make improvements to the procedure employed.

Experiment 2 concerned the training of relational concepts that bear no simple relationship to the experiments or theories considered in the current article. Briefly, each presented stimulus was two size-brightness squares, and for each presented stimulus those two squares were identical on one dimension and either slightly, or greatly, different on the other. "Dimensional" classification involved making one response to pairs that differed in size, and a different response to pairs that differed in brightness. "Similarity" classification involved making one response to pairs where the difference was large, and a different response to pairs where the difference was small. Both relations seem quite derived, and it is not immediately clear to us what either Differentiation Theory or Combination Theory would predict under such circumstances.

Experiment 3 of Smith and Kemler Nelson (1984) was the focus of Experiment 2 in the current paper. Experiment 4 used the pair stimulus type of Experiment 2 of Smith and Kemler Nelson, and did not manipulate time pressure (there was only a speeded condition). Experiments 5 and 6 employed the standard triad procedure, with Experiment 5 employing an instructional manipulation, and Experiment 6 employing concurrent load. One possibility for future research would be to re-examine these manipulations using response-set analysis. Our prediction is that, like our re-examination of Smith and Kemler Nelson's time pressure manipulation, such a re-examination would support Combination Theory over Differentiation Theory. We make this prediction on the basis that the triad procedure, plus response-set analysis, seems to lead to largely the same conclusions as the match-to-standards procedure (which uses a response-set analysis in any case). Instructional manipulations and concurrent load support Combination Theory over Differentiation Theory within the match-to-standards procedure (Wills, Milton, Longmore, Hester & Robinson, 2013).

4.3. Kemler Nelson (1984)

Experiment 1 of Kemler Nelson (1984) was the basis of Experiments 4A and 4B in the current study. Although Kemler Nelson's paper included three different incidental vs. intentional training experiments, we chose Experiment 1 because of the previous (and, as it turns out, incorrect) consensus that the results of Experiment 1 could not be explained through an increase in non-criterial single-attribute responding under incidental conditions (Kemler Nelson, 1984; Ward & Scott, 1987; Ward, 1988). But what of the other experiments in Kemler Nelson (1984)?

Experiment 2 of Kemler Nelson (1984) was similar to Experiment 1 in that it compared intentional training with incidental training. However, incidental training was instantiated in quite a different way to Experiment 1. Specifically, in each trial of the incidental condition of Experiment 2, participants were presented, sequentially, with two faces from the same category (e.g. two policemen) and asked to judge which of the two faces best matched their pre-experimental stereotype (e.g. "which looks more like a policeman?"). This procedure seems to have a number of disadvantages compared to the Experiment 1 procedure. First, this "incidental" procedure explicitly and repeatedly draws participants' attention to the category membership of the items, and that category membership information is vital to answering the question posed by the experimenter in the incidental training phase. Although the term "incidental" is somewhat vague, we feel that this procedure is an atypical example of the concept of incidental category learning. In addition, the Experiment 2 incidental training procedure confounds a number of other variables with the intentional vs. incidental manipulation. For example, the criterial attribute can *never* inform a decision about which of the two same-category faces are more like a policeman (or fireman), because, by definition, all members of a category have the criterial attribute in common. The Experiment 2 incidental training procedure may therefore have the effect of directing attention away from the criterial attribute.[1] The incidental condition also involves a systematically different order of presentation of training stimuli to that in the intentional condition (e.g. the presentation sequence policeman-fireman-policeman can never occur in the incidental condition). In summary, the "incidental" training condition in Experiment 2 does not seem particularly incidental, and also has a number of confounds not present in the incidental training condition of Experiment 1.

Nevertheless, one can still examine the effects of this sort-of-incidental procedure on the prevalence of Overall Similarity, Criterial Attribute, and Non-criterial Attribute responding. Kemler Nelson's original presentation of these data considers only the critical test stimuli and thus cannot distinguish Overall Similarity responding from Non-criterial Attribute responding. However, Kemler Nelson (1988) presented a re-analysis of these data that, although falling short of a full response-set analysis, did permit some assessment of the role of Non-criterial Attribute responding.

Kemler Nelson reported that the intentional versus incidental manipulation did not affect the prevalence of Non-criterial Attribute responding, and significantly increased the prevalence of Overall Similarity responding. From the data reported by Kemler Nelson (1988) it is also possible to determine that, under Kemler Nelson's analysis, Criterial Attribute responding was significantly lower under incidental training than under intentional training, $\chi^2(1) = 8.00, p = .004$. In summary, then, the results of Kemler Nelson (1984, Experiment 2) appear to be that incidental training reduces the prevalence of Criterial Attribute responding and increases the prevalence of Overall Similarity responding. As stated by Kemler Nelson (1988) this finding is consistent with Differentiation Theory. However, this pattern of results is also consistent with Combination Theory. Combination Theory makes no necessary

[1] This point was originally made by Ward and Scott (1987)

predictions about the relative effort involved in Overall Similarity versus Criterial Attribute responding. Overall Similarity responding is more effortful than responding on the basis of a single attribute under this theory, but discovering the criterial attribute also requires effort. Whether Overall Similarity responding or Criterial Attribute responding is the more effortful overall depends on whether discovering the criterial attribute is more effortful than responding on the basis of overall similarity. Thus the results of Kemler Nelson (1984, Experiment 2), like the results in the current paper, do not favor Differentiation Theory over Combination Theory.

This conclusion also potentially provides a resolution to a long-standing disagreement about Experiment 2 of Kemler Nelson (1984). Ward and Scott (1987) replicated Kemler Nelson (1984, Experiment 2) and applied the same semi-response-set analysis applied by Kemler Nelson (1988). However, Ward and Scott concluded that the effect of incidental training was to decrease the prevalence of Criterial Attribute responding and increase the prevalence of Non-criterial attribute responding; no effect on the prevalence of Overall Similarity responding was observed in Ward and Scott's replication of Kemler Nelson (1984, Experiment 2). Ward and Scott's results were interpreted by both Ward and Kemler Nelson as supporting Combination Theory over Differentiation Theory, and thus being contradictory to Kemler Nelson's results (which both parties considered to support Differentiation Theory over Combination Theory). This discrepancy between rather similar studies led to an apparently unresolved disagreement between the authors about the causes of the differences. Kemler Nelson (1988) argued that the difference was due to the fact Ward and Scott adopted a lower learning criterion than in the original study. Ward (1988) rejected this explanation and suggested that the difference lay in a minor procedural difference (Ward and Scott presented the same-category stimulus pairs simultaneously, while Kemler Nelson presented them sequentially). We offer the resolution that the results of Ward and Scott (1987, Experiment 3)— increased Non-criterial Attribute responding combined with reduced Criterial Attribute responding—are consistent with both Combination Theory and Differentiation Theory (see main paper). Note that the same is also true of Kemler Nelson (1984, Experiment 2)—the results of this experiment are consistent with both Combination Theory and Differentiation Theory. There is thus no major discrepancy between the two studies; both provide evidence that is consistent with both theories. The main difference between the two studies is simply in the overall prevalence of Overall Similarity responding (higher in Kemler Nelson than in Ward and Scott). The procedural difference highlighted by Ward, and the procedural difference highlighted by Kemler Nelson, could both potentially account for this difference.

Experiment 3 of Kemler Nelson (1984) also involves an "incidental" training procedure where the participants' attention is explicitly drawn to the category membership of each item, and this information is required to answer the questions posed during incidental training (which involve rating each face on the extent to which it conforms to a pre-experimental stereotype). When the sort-of-incidentally trained participants are later transferred to an intentional training procedure, they outperform participants who had been trained on the intentional procedure from the outset. Further, the advantage of incidental training is larger

for a nonlinearly separable category structure than it is for a linearly separable category structure. One explanation for this result, similar to one of the explanations offered by Kemler Nelson, is that this "incidental" training procedure promotes a strategy of exemplar memorization. Formal modeling demonstrates that exemplar memorization predicts better performance on the nonlinear structure than on the linear structure (Medin & Schwanenflugel, 1981), and this is the result observed in Kemler Nelson's incidental condition. In the intentional condition, there is no significant difference in the rate at which the linear and nonlinear structures are learned. This lack of difference might reflect a mixture of exemplar memorization and other (e.g. dimensional summation) strategies, although note that advantages of nonlinear over linear structures are also sometimes seen under intentional conditions (Medin & Schwanenflugel, 1981). Although this is an interesting hypothesis, a strategy of exemplar memorization is consistent with both Combination Theory and Differentiation Theory. It is true that Combination Theory predicts that an exemplar memorization strategy would not be possible if time and cognitive resources were sufficiently limited (because exemplar memorization would require the combination of attributes). However, as the test phase involves intentional training in both conditions of Kemler Nelson's Experiment 3, there is no reason to assume time or cognitive resources were particularly limited in this case.

Experiment 4 of Kemler Nelson (1984) is developmental in nature. The central result is that Kindergarten children find that a category structure with one criterial attribute and three entirely nondiagnostic attributes (CA structure) is harder to learn than a category structure with four .75 characteristic attributes (FR structure). Fifth graders are better on both problems than Kindergarden children; they still find the CA structure harder, but the improvement on the harder CA structure is greater. The result that the CA structure is harder to learn is unsurprising from the perspective of Combination Theory - whether the FR or CA structure is harder to learn depends on whether discovering the criterial attribute is more effortful than the combination required for overall similarity classification. Kemler Nelson's data and our own suggest that discovering the criterial attribute is more effortful. The idea that cognitive development improves performance more on hard problems than easy problems seems possible to accommodate from a number of different theoretical perspectives. Combination Theory, for example, might propose that searching for a criterial attribute is something that children develop, or are taught, during their primary school years. In a more recent study, Minda & Miles (2009) compared adults and children on the criterial-attribtue procedure. An overall similarity strategy was rare in both groups, and Criterial Attribute responding was more common in adults than in children. These results are consistent with the conclusions of the current paper.

4.4. Smith and Shapiro (1989)

Smith and Shapiro (1989) reported three experiments. Experiment 1 was the focus of our investigation in Experiment 5. Experiment 2 employed the same stimulus set (pronounceable nonwords with a criterial attribute structure), and had three conditions - intentional unspeeded training (10 seconds per stimulus), intentional speeded training (1 second per stimulus), and incidental training. The time pressure manipulation had no significant effect, and is not considered further. Incidental training involved asking the participants to read

the pronounceable nonwords aloud as quickly as possible; category information was available in the sense that Group 1 words appeared on the left and Group 2 words appeared on the right, but the relevance of this positioning was not revealed until the test phase. Thus, Experiment 2, like Experiment 1 of Kemler Nelson (1984), but unlike the later experiments in that paper, seems to fit typical conceptions of what it means to train incidentally. In Smith and Shapiro (1989), Experiment 2, the effect of incidental training was to increase the proportion of "true" family resemblance responders, relative to both the unspeeded and speeded intentional training conditions. This conclusion is based on the same analysis in their Experiment 1, and thus is open to the same criticism that overall similarity responding is not fully distinguished from non-criterial attribute responding in this analysis (see the main body of the current paper for details).

It is possible to calculate from figures given by Smith and Shapiro that the mean consistency of single-attribute models in the incidental condition of Experiment 2 would have been .85. Although it is not possible to come to definitive conclusions without further details that Smith and Shapiro do not provide, and although the data they do provide combine criterial-attribute responding with non-criterial attribute responding, a mean consistency of .85 does give at least some indication that single-attribute strategies do a reasonably good job of describing behavior in the incidental condition of Smith and Shapiro (1989, Experiment 2). One possibility for future research would be to conduct a large-scale replication of this experiment and apply a full response-set analysis of the data thus produced. On the basis of Experiments 4A and 4B in the current paper, which differ from Smith and Shapiro (1989, Experiment 2) only in the details of how the abstract structure is instantiated with physical stimuli and how incidental and intentional training is achieved, our prediction is that such a replication would not favor Differentiation Theory over Combination Theory.

Experiment 3 of Smith and Shapiro (1989) was a further investigation of the effects of time pressure on intentional training within the criterial-attribute procedure. Relative to Experiment 2, the main difference was that the time pressure was more extreme (6 seconds in total for 24 training stimuli). Time pressure *reduces* the proportion of "true FR" responders in this Experiment, a result which, on the face of it, seems more consistent with Combination Theory than Differentiation Theory. Smith and Shapiro report no inferential statistics for this experiment, but nevertheless argue that the results of their time pressure experiments are at variance with both the other results in their paper, and the results of both Kemler Nelson (1984), and Smith and Kemler Nelson (1984). Their explanation of this difference centers around an argument that time pressure has opposite effects on spontaneous classification (Smith & Kemler Nelson, 1984) and category learning (Kemler Nelson, 1984; Smith & Shapiro, 1989), while concurrent load has the same effect on spontaneous classification and category learning. The work in the current paper suggests an alternative explanation—the effects of time pressure and concurrent load on both spontaneous classification and category learning are basically the same; they both increase the prevalence of classifying on the basis of a single, arbitrarily chosen, dimension.

Appendix: By-experiment analyses of consistency and margin

In the following sections, we examine consistency and margin for each model in each condition of each experiment. In some cases, this results in small sample sizes; inferential statistics are only reported in cases where N is at least 10. Raw data and analysis scripts for these analyses are available as part of the raw data archives for Experiments 1, 2, 3A, 3B, 4A, 4B and 5 (see main paper).

Condition	Measure	UD	OS	ID	*All models*
640 ms	**prop.**	**.74**	**.22**	**.04**	
	consist.	.66	.53	.48	.62
	margin	.19	.10	.02	.17
	N	20	6	1	
1024 ms	**prop.**	**.76**	**.24**	**.00**	
	consist.	.70	.57	–	.67
	margin	.21	.12	–	.19
	N	22	7	0	
2048 ms	**prop.**	**.76**	**.10**	**.14**	
	consist.	.71	.63	.80	.71
	margin	.20	.08	.26	.19
	N	22	3	4	
3072 ms	**prop.**	**.44**	**.20**	**.36**	
	consist.	.75	.63	.84	.75
	margin	.23	.14	.32	.24
	N	11	5	9	
7500 ms	**prop.**	**.28**	**.36**	**.36**	
	consist.	.83	.64	.95	.81
	margin	.27	.15	.44	.29
	N	8	10	10	
All conditions	consist.	.71	.60	.86	
	margin	.21	.13	.35	

Table 4: Extended response-set analyses of Experiment 1. Prop.: Proportion of participants best fit by a Unidimensional (UD), Overall Similarity (OS), and Identity (ID) response model. Consist.: mean proportion of responses predicted by the best-fitting model. Margin: mean difference in proportion of responses predicted by the best-fitting and second-best-fitting model. N: sample size

Experiment 1

Considering the "All conditions" section of the Table 4 first, which collapses across experimental condition, it is apparent that all three strategies show at least reasonable consistency. It's also the case that the three strategies differ in mean consistency, $F(2, 135) = 32.79, p < .0001$, with all pairwise comparisons significant according to a Tukey HSD test,

max. $p < .0001$. Similar conclusions can be drawn from the *margin* variable. Specifically, the additional proportion of responses predicted by the winning model, relative to its nearest competitor, is fairly substantial for all models, and the three strategies differ in mean margin, $F(2, 135) = 22.07, p < .0001$, with all pairwise comparisons significant according to a Tukey HSD test, max. $p < .005$. The Overall Similarity strategy has the lowest mean consistency and lowest mean margin but, as we shall see, this is not invariably the case across experiments.

Turning to the effects of time pressure (the "all models" section of the Table), mean consistency decreases with increasing time pressure, $r^2_{adj} = .77, p = .03$, as does mean margin, $r^2_{adj} = .88, p = .01$. A plausible explanation is that time pressure reduces the extent to which participants can apply their intended strategy consistently, affecting both model consistency and the extent to which one model is a clear winner (although note that both mean consistency and mean margin remain substantial, even at a 640 ms presentation time). Statistical investigation of the interaction between time pressure and response strategy would be largely uninformative for the current data set due to the small sample sizes in many cells ($N < 10$ for 9 of the 16 cells).

Condition	Measure	UD	OS	ID	*All models*
1000 ms	**prop.**	**.38**	**.13**	**.49**	
	consist.	.63	.55	.78	.69
	margin	.13	.09	.24	.18
	N	14	5	18	
5000 ms	**prop.**	**.10**	**.00**	**.90**	
	consist.	.76	–	.89	.88
	margin	.16	–	.37	.35
	N	4	0	37	
All conditions	consist.	.66	–	.85	
	margin	.14	–	.33	

Table 5: Extended response-set analyses of Experiment 2. Prop.: Proportion of participants best fit by a Unidimensional (UD), Overall Similarity (OS), and Identity (ID) response model. Consist.: mean proportion of responses predicted by the best-fitting model. Margin: mean difference in proportion of responses predicted by the best-fitting and second-best-fitting model. *N*: sample size

Experiment 2

Only 5 people in total were best fit by the OS response model (see Table 5). This sample is too small for meaningful analysis of consistency or margin. The UD and ID response models show a similar pattern of results to that observed in Experiment 1. Both models show at least reasonable consistency and margin. Consistency is higher for the ID model than the UD model, $t(71) = 5.41, p < .0001$, as is margin, $t(71) = 4.84, p < .0001$. Time pressure reduces consistency, $t(76) = 6.11, p < .0001$, and margin, $t(76) = 5.19, p < .0001$. Note that the ID model is more prevalent in the low time pressure condition, so it is not

clear which of these effects is cause and which is effect—the effects of time pressure on model consistency could be due to the difference in model consistency or vice versa. The sample size for the UD model in the 5000 ms condition is too small to permit meaningful analysis of the interaction between time pressure and model type.

Condition	Measure	UD	OS	ID	*All models*
2000 ms	**prop.**	**.29**	**.67**	**.04**	
	consist.	.70	.75	.70	.73
	margin	.13	.24	.16	.21
	N	16	36	2	
5000 ms	**prop.**	**.40**	**.42**	**.18**	
	consist.	.75	.77	.75	.76
	margin	.18	.26	.21	.22
	N	20	21	9	
All conditions	consist.	.73	.76	.74	
	margin	.16	.25	.20	

Table 6: Further response-set analyses of Experiment 3A. Prop.: Proportion of participants best fit by a Unidimensional (UD), Overall Similarity (OS), and Identity (ID) response model. Consist.: mean proportion of responses predicted by the best-fitting model.Margin: mean difference in proportion of responses predicted by the best-fitting and second-best-fitting model. N: sample size

Experiment 3A

Although all models show good consistency and margin, Table 6 reveals a slightly different pattern of results in Experiment 3A, compared to Experiments 1 and 2. In the current experiment UD, OS and ID models do not differ significantly in consistency, $F(2, 101) < 1$ and, while margin does differ significantly, $F(2, 101) = 4.81, p = .01$, it is the lower margin of the UD model relative to the OS model that drives this effect, $p < .01$, with the other two pairwise comparisons not approaching significance on a Tukey HSD, $p > .5$. Thus in Experiment 3A, where the OS model provides the best fit for approximately half the participants, there is no evidence that the OS model is a worse description of peoples' behavior than the other two models (or, alternatively, no evidence that OS responders find it no more difficult to apply their strategy consistently than do UD or ID responders). Time pressure has no significant effect on consistency, $t(102) = 1.15, p = .25$, or on margin, $t(102) < 1$, in Experiment 3A. Interactions between model type (UD vs. OS) and time pressure on the effects of consistency and margin can be meaningfully examined in this data set (ID must be excluded due to low sample size). Neither interaction is significant, $F < 1$ in both cases.

Experiment 3B

Consistency and margin in Experiment 3B are similar to Experiment 3A (see Table 7). Collapsing across conditions, all three models have at least reasonable consistency and

Condition	Measure	UD	OS	ID	*All models*
< 2000 ms	**prop.**	**.09**	**.88**	**.03**	
	consist.	.68	.69	.38	.68
	margin	.16	.23	.02	.21
	N	3	28	1	
> 5000 ms	**prop.**	**.21**	**.56**	**.22**	
	consist.	.68	.80	.81	.77
	margin	.14	.26	.29	.24
	N	7	18	7	
All conditions	consist.	.68	.73	.76	
	margin	.15	.24	.26	

Table 7: Further response-set analyses of Experiment 3B. Prop.: Proportion of participants best fit by a Unidimensional (UD), Overall Similarity (OS), and Identity (ID) response model. Consist.: mean proportion of responses predicted by the best-fitting model. Margin: mean difference in proportion of responses predicted by the best-fitting and second-best-fitting model. N: sample size

margin. The sample size for the ID model is a little low to draw conclusions. The UD and OS models do not differ significantly in consistency, $t(54) = 1.25, p = .22$, but the OS model has a significantly larger margin, $t(54) = 2.14, p = .04$. Hence, in Experiment 3B, where around 70% of participants are classified as OS responders, the OS response model wins more decisively than the UD model (the opposite pattern of result to experiments where the UD model dominates). Across experiments, a picture emerges that the UD and OS response models fare better or worse in different experiments in terms of capturing individual's performances, but one is not uniformly superior to another.

In Experiment 3B, time pressure significantly reduces consistency, $t(62) = 3.13, p = .003$, but not margin, $t(62) < 1$. Cell sample sizes are too small for meaningful interaction analysis.

Experiment 4A

Considering the primary analysis first (i.e. analysis of participants passing criterion, using all test phase stimuli in the analysis), only three participants in total were best fit by the OS response model (see Table 8, "passed, all-items" rows). Although this sample is too small for meaningful analysis of consistency or margin, we note in passing that the margin for the OS response model appears to be very small, with zero margin for the one participant best fit by the OS model in the incidental condition. Zero margin represents a tie; the OS model tied with the NCA model for this participant. The tie was awarded to the OS model in deference to the traditional view that OS responding is common under incidental training conditions. No other participant had a zero margin.

Turning to the CA and NCA models, both show good consistency and margin. Consistency is higher for the CA model than the NCA model, $t(45) = 3.23, p = .002$, as is the margin, $t(45) = 2.73, p = .009$. Incidental training reduces consistency, relative to intentional training, $t(48) = 3.18, p = .003$; it also reduces margin, $t(48) = 2.38, p = .02$. Note that

	Condition	Measure	OS	CA	NCA	*All models*
Passed, all-items	Intentional	**prop.**	**.07**	**.68**	**.25**	
		consist.	.71	.96	.83	.92
		margin	.01	.28	.18	.24
		N	2	19	5	
	Incidental	**prop.**	**.03**	**.44**	**.53**	
		consist.	.72	.84	.79	.82
		margin	.00	.19	.16	.17
		N	1	13	10	
	Both	consist.	.72	.92	.81	
		margin	.01	.24	.17	
ALL, all-items	Intentional	**prop.**	**.06**	**.66**	**.28**	
		consist.	.71	.93	.78	.88
		margin	.01	.26	.15	.22
		N	2	21	7	
	Incidental	**prop.**	**.01**	**.34**	**.65**	
		consist.	.72	.73	.71	.72
		margin	.00	.14	.15	.14
		N	1	22	26	
	Both	consist.	.72	.83	.73	
		margin	.01	.20	.15	
ALL, 10-items	Intentional	**prop.**	**.12**	**.66**	**.22**	
		consist.	.84	.93	.81	.90
		margin	.04	.17	.13	.14
		N	4	21	6	
	Incidental	**prop.**	**.11**	**.27**	**.62**	
		consist.	.72	.84	.73	.76
		margin	.02	.12	.11	.10
		N	6	15	23	
	Both	consist.	.77	.89	.74	
		margin	.03	.14	.11	

Table 8: Further response-set analyses of Experiment 4A. Prop.: Proportion of participants best fit by an Overall Similarity (OS), Criterial Attribute (CA) and Non-criterial Attribute (NCA) response model. Consist.: mean proportion of responses predicted by the best-fitting model. Margin: mean difference in proportion of responses predicted by the best-fitting and second-best-fitting model. *N*: sample size

the CA model is more prevalent in the intentional condition, where consistency is higher, so it is not clear which of these phenomena is cause and which is effect. Sample size for the NCA model in the intentional condition is too small to permit meaningful analysis of the interaction between training type and model type.

The "ALL, all-items" rows of Table 8 present the above all-test-items analysis, but with all participants included. Consistencies and margins are slightly lower overall, which is to be expected given the sample now includes participants who did not meet the learning criterion. The central conclusions are unchanged. Consistency is higher for the CA model than the NCA model, $t(74) = 2.82, p = .006$, as is the margin, $t(74) = 2.26, p = .03$. Incidental training reduces consistency, relative to intentional training, $t(77) = 4.88, p < .0001$; it also reduces margin, $t(77) = 3.36, p = .001$.

Turning to the all-participants analysis excluding OS-model-ambiguous test trials (the "ALL, 10-items" rows of Table 8), there are now ten participants in total who are best fit by the the OS model. There is a significant effect on model type on consistency, $F(2, 72) = 11.07, p < .0001$, with the difference between the NCA and OS models not approaching significance on a Tukey HSD test, $p = .86$. The other two pairwise differences were significant, $p < .05$. The OS model still has a small margin, both in absolute terms and relative to the NCA and CA models, $F(2, 72) = 15.29, p < .0001$, CA-OS $p < .0001$, NCA-OS $p < .001$, NCA-CA $p = .08$.

Experiment 4B

The results for Experiment 4B were similar to those of Experiment 4A. Considering the primary analysis first (Table 9, "passed" rows), seven participants were best fit by the OS response mode; too small a sample for meaningful analysis. Turning to the CA and NCA models, both show good consistency and margin. Consistency is higher for the CA model than the NCA model, $t(42) = 2.63, p = .01$, as is the margin, $t(42) = 2.97, p = .005$. Incidental training reduces consistency, relative to intentional training, $t(49) = 4.92, p < .0001$; it also reduces margin, $t(49) = 4.22, p = .0001$.

The analysis considering all participants is presented in the "ALL" rows of Table 9. Inclusion of all participants makes very little difference. The sample size for the OS model is still too small for meaningful analysis. The CA and NCA models both show good consistency and margin. Consistency is higher for the CA model than the NCA model, $t(49) = 2.81, p = .007$, as is the margin, $t(49) = 3.07, p = .004$. Incidental training reduces consistency, relative to intentional training, $t(57) = 6.00, p < .0001$; it also reduces margin, $t(57) = 4.69, p < .0001$. One slightly counter-intuitive aspect of the "all-participants analysis" is that the sample sizes for the intentional condition are identical to the previous analysis, despite some participants in this condition not meeting the learning criterion. This is because the three participants excluded by the learning criterion were best fit by a response bias model (pressing the same key on all trials) and hence were excluded from this analysis also.

Experiment 5

The results for Experiment 5 were, in most respects, similar to those of Experiments 4A and 4B. In the "passed, all-items" analysis (see Table 10), no participants were best fit

	Condition	Measure	OS	CA	NCA	*All models*
Passed	Intentional	**prop.**	**.09**	**.63**	**.28**	
		consist.	.88	.98	.89	.95
		margin	.08	.19	.14	.17
		N	3	20	5	
	Incidental	**prop.**	**.15**	**.33**	**.52**	
		consist.	.71	.85	.83	.81
		margin	.01	.12	.11	.10
		N	4	9	10	
	Both	consist.	.78	.94	.85	
		margin	.04	.17	.12	
ALL	Intentional	**prop.**	**.09**	**.63**	**.28**	
		consist.	.88	.98	.89	.95
		margin	.08	.19	.14	.17
		N	3	20	5	
	Incidental	**prop.**	**.15**	**.33**	**.52**	
		consist.	.68	.82	.81	.79
		margin	.01	.12	.11	.09
		N	5	11	15	
	Both	consist.	.76	.92	.83	
		margin	.03	.16	.11	

Table 9: Extended response-set analyses of Experiment 4B. Prop.: Proportion of participants best fit by an Overall Similarity (OS), Criterial Attribute (CA) and Non-criterial Attribute (NCA) response model. Consist.: mean proportion of responses predicted by the best-fitting model. Margin: mean difference in proportion of responses predicted by the best-fitting and second-best-fitting model. *N*: sample size

by the OS model. Both the CA and NCA models have very good consistency and margin. Consistency is higher for the CA model than the NCA model, $t(49) = 2.53, p = .01$, as is margin, $t(49) = 2.23, p = .03$. Interestingly, concurrent load did not significantly affect consistency, $t(49) = 1.12, p = .27$, or margin, $t(49) < 1$. Experiment 5 thus provides some evidence that the CA model genuinely has higher consistency than the NCA model (this model-type effect being confounded with the effect of training type in Experiments 4A and 4B). Whether this difference reflects a difference in the quality of these accounts of behaviour, or whether it reflects the fact that participants applying a NCA strategy find it hard to do so consistently, is a matter for future research.

Turning to the all-participants, all-test-items analysis, again no participants were best fit by the OS model. The CA and NCA models have good consistency and margin, albeit a little lower than when participants failing the learning criterion were excluded. Consistency is higher for the CA model than the NCA model, $t(59) = 3.74, p = .0004$, as is margin, $t(59) = 3.31, p = .002$. In this analysis considering all participants, concurrent load reduces consistency, relative to full-attention conditions, $t(59) = 3.17, p = .001$; it also reduces

	Condition	Measure	OS	CA	NCA	*All models*
Passed, all-items	Concurrent load	**prop.**	**.00**	**.48**	**.52**	
		consist.	–	.93	.92	.92
		margin	–	.25	.25	.25
		N	0	10	7	
	Full attention	**prop.**	**.00**	**.74**	**.26**	
		consist.	–	.97	.87	.95
		margin	–	.28	.19	.27
		N	0	28	6	
	Both	consist.	–	.96	.89	
		margin	–	.27	.22	
ALL, all-items	Concurrent load	**prop.**	**.00**	**.41**	**.59**	
		consist.	–	.86	.78	.82
		margin	–	.22	.18	.20
		N	0	13	12	
	Full attention	**prop.**	**.00**	**.68**	**.32**	
		consist.	–	.97	.82	.93
		margin	–	.28	.19	.26
		N	0	28	8	
	Both	consist.	–	.93	.79	
		margin	–	.26	.19	
ALL, 10-items	Concurrent load	**prop.**	**.00**	**.46**	**.54**	
		consist.	–	.89	.87	.88
		margin	–	.22	.15	.14
		N	0	12	9	
	Full attention	**prop.**	**.02**	**.66**	**.32**	
		consist.	.86	.97	.87	.94
		margin	.03	.18	.10	.16
		N	1	27	7	
	Both	consist.	.86	.95	.87	
		margin	.03	.17	.12	

Table 10: Further response-set analyses of Experiment 5 Prop.: Proportion of participants best fit by an Overall Similarity (OS), Criterial Attribute (CA) and Non-criterial Attribute (NCA) response model. Consist.: mean proportion of responses predicted by the best-fitting model. Margin: mean difference in proportion of responses predicted by the best-fitting and second-best-fitting model. *N*: sample size

margin, $t(59) = 2.79, p = .007$. One explanation of the apparent difference between the current analysis and the one above is that those participants who failed the learning criterion were particularly distracted by the concurrent load task.

The conclusions regarding Experiment 5 are unchanged if the OS-ambiguous trials are ignored (the "ALL, 10-items" rows of Table 10). Only one participant is best fit by the OS model; no conclusions concerning consistency and margin can be drawn from this single data point. The CA and NCA models continue to have good consistency and margin in this analysis. Consistency is higher for the CA model than the NCA model, $t(53) = 2.66, p = .01$, as is margin, $t(53) = 3.53, p = .001$. Concurrent load decreases consistency relative to full attention, $t(54) = 2.65, p = .01$; the corresponding decrease in margin is not significant, $t(54) = 1.26, p = .21$.

References

Block, J., Block, J., & Harrington, D. (1974). Some misgivings about the matching familiar figures test as a measure of reflection-impulsivity. *Developmental Psychology*, *10*, 611–632.

Edmunds, C., Milton, F., & Wills, A. (accepted). Feedback can be superior to observational training for both rule-based and information-integration category structures. *Quarterly Journal of Experimental Psychology*, .

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*, 27–38.

Helmers, K., Young, S., & Pihl, R. (1995). Assessment of measures of impulsivity in healthy male volunteers. *Personality and Individual Differences*, *19*, 927–935.

Henize, G., Ploner, M., Dunkler, D., & Southworh, H. (2013). *logistf: Firth's bias reduced logistic regression*. URL: `http://CRAN.R-project.org/package=logistf`.

Kagan, J. (1965). Impulsive and reflective children: Significance of conceptual tempo. In J. Krumboltz (Ed.), *Learning and the educational process*. Chicago: Rand McNally.

Kemler Nelson, D. G. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior*, *23*, 734–759.

Kemler Nelson, D. G. (1988). When category learning is holistic: A reply to Ward and Scott. *Memory & Cognition*, *16*, 79–89.

Lea, S. E. G., WIlls, A. J., Leaver, L. A., Ryan, C. M. E., Bryant, C. M. L., & Millar, L. (2009). A comparative analysis of the categorization of multidimensional stimuli: II. Strategic information search in humans (Homo sapiens) but not in pigeons (Columba livia). *Journal of Comparative Psychology*, *123*, 406–420.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 355–368.

Minda, J. P., & Miles, S. J. (2009). Learning new categories: Adults tend to use rules while children sometimes rely on family resemblance. In N. A. Taatgen, & H. van Rihn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1518–1523). Austin, TX: Cognitive Science Society.

Patton, J., Stanford, M., & Barratt, E. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, *51*, 768–774.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: `http://www.R-project.org/`.

Raijmakers, M. E. J., Jansen, B. R. J., & van der Maas, H. L. J. (2004). Rules and development in triad classification task performance. *Developmental Review*, *24*, 289–321.

Smith, J. D., & Kemler Nelson, D. G. (1984). Overall similarity in adults' classification: The child in all of us. *Journal of Experimental Psychology: General*, *113*, 137–159.

Smith, J. D., & Shapiro, J. H. (1989). The occurrence of holistic categorization. *Journal of Memory and Language*, *28*, 386–399.

Stanford, M. S., Mathias, C. W., Dougherty, D. M., Lake, S. L., Anderson, N. E., & Patton, J. H. (2009). Fifty years of the Barratt Impulsiveness Scale: An update and review. *Personality and Individual Differences*, *47*, 385–395.

Thompson, L. A. (1994). Dimensional strategies dominate perceptual classification. *Child Development*, *65*, 1627–1645.

Verbruggen, F., Adams, R., & Chambers, C. D. (2012). Proactive motor control reduces monetary risk taking in gambling. *Psychological Science*, *23*, 805–815.

Ward, T. B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 103–112.

Ward, T. B. (1988). When is category learning holistic? A reply to Kemler Nelson. *Memory & Cognition*, *16*, 85–89.

Ward, T. B., & Scott, J. (1987). Analytic and holistic modes of learning family-resemblance concepts. *Memory and Cognition*, *15*, 42–54.

Wills, A. J., Milton, F., Longmore, C. A., Hester, S., & Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification? *Quarterly Journal of Experimental Psychology*, *66*, 299–318.