Combination or Differentiation? Two theories of processing order in classification

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

[a] *School of Psychology, Plymouth University, UK. Tel: +44 7540 754852*
[b] *Psychology, University of Exeter, UK*

Abstract

Does cognition begin with an undifferentiated stimulus whole, which can be divided into distinct attributes if time and cognitive resources allow (Differentiation Theory)? Or does it begin with the attributes, which are combined if time and cognitive resources allow (Combination Theory)? Across psychology, use of the terms analytic and non-analytic imply that Differentiation Theory is correct — if cognition begins with the attributes, then synthesis, rather than analysis, is the more appropriate chemical analogy. We re-examined four classic studies of the effects of time pressure, incidental training, and concurrent load on classification and category learning (Kemler Nelson, 1984; Smith & Kemler Nelson, 1984; Smith & Shapiro, 1989; Ward, 1983). These studies are typically interpreted as supporting Differentiation Theory over Combination Theory, while more recent work in classification (Milton, Longmore & Wills, 2008, *et seq.*) supports the opposite conclusion. Across seven experiments, replication and re-analysis of the four classic studies revealed that they do not support Differentiation Theory over Combination Theory — two experiments support Combination Theory over Differentiation Theory, and the remainder are compatible with both accounts. We conclude that Combination Theory provides a parsimonious account of both classic and more recent work in this area. The presented data do not require Differentiation Theory, nor a Combination-Differentiation hybrid account.

Keywords: categorization, category learning, time pressure, concurrent load, holistic, analytic, nonanalytic

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

"Synthesis of any particular letter or figure takes an appreciable time" (Neisser, 1967, p. 103)

1. Introduction

In this article we consider two, approximately opposite, theories of how the psychological processes underlying a classification decision unfold as more time or cognitive resources become available. These two theories might be more accurately described as frameworks, as each captures the basic operating principles of a class of specific theories that are nonidentical. However, this classification of theories is a relatively natural one, with substantial within-category similarities and between-category differences.

The two theories are described here as Combination Theory and Differentiation Theory. The current usage of these terms is novel, but they are intended to capture ideas already present in the literature. Differentiation Theory assumes that classification starts with an undifferentiated whole (a "holistic blob", Lockhead, 1972), which can be broken into its constituent attributes if time and cognitive resources allow. In contrast, Combination Theory assumes classification starts with the attributes, and that information from these attributes can be combined and weighted if time and cognitive resources allow. The widely-used terms "analytic" and "nonanalytic" (Brooks, 1978) presuppose Differentiation Theory—as we will discuss below, they make little sense if Combination Theory is correct.

The question of whether Differentiation Theory or Combination Theory provides the better explanation of classification is controversial. On the one hand, a series of classic studies (Kemler Nelson, 1984; Smith & Kemler Nelson, 1984; Smith & Shapiro, 1989; Ward, 1983) are typically considered to support Differentiation Theory over Combination Theory (e.g. Couchman, Coutinho & Smith, 2010; Goldstone & Barsalou, 1998). On the other hand, a series of more recent studies employing a slightly different procedure (Milton, Longmore & Wills, 2008; Milton, Wills & Hodgson, 2009; Wills, Milton, Longmore, Hester & Robinson, 2013b), largely support the opposite conclusion. The current investigation offers a reconciliation of these apparently incompatible studies. The reconciliation we offer is that Combination Theory provides a parsimonious account of both sets of studies. Differentiation Theory is directly disconfirmed in two cases.

1.1. Combination Theory

In Combination Theory, the input to the classification system is in the form of a set of distinct attributes (dimensions or features). Classification on the basis of multiple attributes involves the collection of information across those attributes, which takes time. Classification on the basis of multiple attributes sometimes involves weighting those attributes differently. This also takes time, possibly more time than employing an unweighted combination.

The intellectual roots of Combination Theory can be traced back at least as far as the fuzzy logical model of perception (Oden & Massaro, 1978; Thompson & Massaro, 1989) and feature-integration theory (Treisman & Gelade, 1980). One formal instantiation of Combination Theory within categorization research is Lamberts's extension to the GCM (Generalized Context Model; Nosofsky, 1984), the EGCM (Extended Generalized Context Model; Lamberts, 1995). EGCM is a stochastic sampling model; each attribute is assigned a hazard function such that the probability of that attribute having been sampled by time zero is zero, increasing thereafter. Thus, the more time available to observe the stimulus, the more dimensions, on average, will be available on which to make a response. Another, less formal, example of Combination Theory is Dimensional Summation theory (Milton & Wills, 2004). Dimensional Summation theory assumes a serial, limited-capacity, rule-like process. Stimulus dimensions are intentionally, sequentially, queried until sufficient information is available to apply the currently selected categorization rule. Taken together, EGCM and Dimensional Summation theory provide two illustrations that Combination Theory can, at the level of detailed process, be implemented in a number of architecturally distinct ways.

1.2. Evidence for Combination Theory

It is perhaps tempting to argue that, from a neuroscientific perspective, Combination Theory has to be correct. Such an argument would point to the fact that, early in the cortical visual processing stream, neurons seem sensitive to the attributes of objects (e.g. Hubel & Wiesel, 1959) while, later in the stream, neurons seem sensitive to specific objects (e.g. Gross, 2008). Such an argument, while intuitively appealing, is not compelling. In the context of classification, the question is not whether combination of attributes occurs, but whether the classification system has direct access to early visual processing. This is a matter of debate (e.g. Pylyshyn, 1999) and it remains entirely possible that the input to the classification system might be in the form of initially

undifferentiated wholes, as assumed by Differentiation Theory (cf. template-matching theories, e.g. Larsen & Bundesen, 1992).

Turning to behavioral data, Combination Theory assumes that classification on the basis of multiple attributes involves the collection of information across those attributes, and that this takes time. Combination Theory therefore predicts Ward and Scott's (1987) observation that participants classifying on the basis of a single attribute have faster reaction times than participants classifying on the basis of overall similarity. Combination Theory also predicts that participants will often choose to classify on the basis of a single dimension, at least in the absence of feedback that indicates this time-efficient strategy is inappropriate (e.g. Medin et al., 1987). Similarly, Combination Theory predicts that single-dimension category structures are generally learned more quickly than multi-attribute category structures (Shepard et al., 1961) — the assumption is that participants test the lowest effort, single-dimension, hypotheses first (cf. Nosofsky et al., 1994). More generally, any manipulation that reduces the time or cognitive resources a person is willing or able to dedicate to the classification decision should reduce the prevalence of multi-attribute classification. This includes reducing the time available for classification, giving the participants a concurrent task, or testing people who are impulsive or have a small working memory capacity.

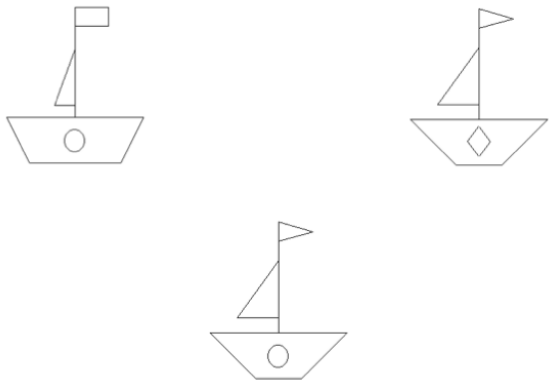

Figure 1: Example trial in a match-to-standards procedure. The top two stimuli are the standards.

The most extended set of experiments in support of Combination Theory employs the match-to-standards procedure (Milton & Wills, 2004; Regehr & Brooks, 1995) and includes all these forms of evidence. Figure 1 illustrates the procedure. In each trial, participants are presented with three stimuli. Two of these stimuli (the standards) are always the same, and these standards differ on all four experimenter-defined binary stimulus dimensions. On most trials, the third stimulus differs by

one attribute from one standard, and by three attributes from the other standard (on the remaining trials, the third stimulus is also one of the two standards). On each trial, the participant is asked which of the two standards the third stimulus "goes with best". This is a spontaneous classification procedure—no feedback on the classification decisions is given. Each participant's set of responses to the stimuli are analysed to determine whether the participant is employing an overall similarity, single-dimension, or other, classification strategy.

Using this match-to-standards procedure, Milton et al. (2008) demonstrated that decreases in stimulus presentation time decrease the prevalence of overall similarity classification (see also Milton, Viika, Henderson & Wills, 2011). Wills, Milton, Longmore, Hester, and Robinson (2013b) further reported that (a) concurrent load decreases the prevalence of overall similarity classification, (b) overall similarity classifiers have larger working memory capacities than single-dimension classifiers, and that (c) instructions to respond meticulously increase the prevalence of overall similarity classification. Wills, Longmore, and Milton (2013a) demonstrated that overall similarity classifiers scored lower on a standard measure of impulsivity than single-attribute classifiers. All these results are consistent with Combination Theory.

The match-to-standards procedure has also been the subject of neuroscientific investigation. Milton, Wills, and Hodgson (2009) reported greater frontal lobe activity in overall similarity classifiers than single-attribute classifiers, supporting the conclusion that overall similarity classification is more effortful. This conclusion is consistent with Combination Theory. In an eye-tracking study, Milton and Wills (2009) reported that participants classifying on the basis of overall similarity produced more fixations overall than participants classifying on the basis of a single dimension, and fixated a greater proportion of the dimensions. Their reaction times were also longer. These results are again consistent with Combination Theory.

In summary, the evidence for Combination Theory is substantial, but the majority of that evidence comes from a single procedure: the match-to-standards procedure. A critic might assert that the match-to-standards procedure is somehow atypical. Responding to this assertion was one of the motivations for the current series of studies.

1.3. Differentiation Theory

In Differentiation Theory, the input to the classification system is in the form of an undifferentiated whole or "blob". The classification system quickly has access to information about the overall

level of similarity between the presented stimulus and other stimuli (presented or remembered), but does not initially have direct access to the different attributes that make up that stimulus. Classification on the basis of a single attribute therefore involves a process of differentiating the input (separating the whole into its constituent parts), and this takes time.

Differentiation Theory has been widely adopted across psychology. Its intellectual roots can be traced back at least as far as Lockhead (1972), and were succinctly summarized by Ward's statement "holistic, integral processing precedes analytic dimensional processing" (Ward, 1983, p. 103), see also Smith and Kemler Nelson (1984). Another intellectual root of Differentiation Theory is the theory of nonanalytic concept formation (Brooks, 1978). Indeed, the now widely-used terms *analytic* and *nonanalytic* presuppose something like Differentiation Theory—categorization can only be "analytic" if there is something to analyze; in other words, something to separate into its constituent elements. From the perspective of Combination Theory, the chemical analogy is synthesis rather than analysis. The continuing influence of Differentiation Theory is also revealed in statements such as "it is easier for people to base similarity and categorization judgments on more, rather than fewer, properties" (Goldstone & Barsalou, 1998, p. 239-240).

One way to conceptualize Differentiation Theory within formal accounts of categorization is to assume that the selective-attention process instantiated in theories such as GCM (Nosofsky, 1984) or ALCOVE (Kruschke, 1992) is effortful and time consuming (Nosofsky & Kruschke, 2002), and that attention is distributed equally across stimulus dimensions prior to this effortful process. One would further have to assume that the stimulus sampling process described in accounts such as EGCM (Lamberts, 1995) took negligible time to complete. Models such as ALCOVE and GCM do not have a sampling process, so they formally instantiate this assumption, although this appears to be a simplification rather than a strong theoretical commitment (Cohen & Nosofsky, 2003). One might also argue that the COVIS model (COmpetition between Verbal and Implicit Systems; Ashby, Alfonso-Reese, Turken & Waldron, 1998) has more in common with Differentiation Theory than with Combination Theory. For example, COVIS predicts that learning of an overall similarity category structure is less affected by time pressure or concurrent load than learning of a single-dimension category structure (Minda et al., 2008; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006).

1.4. Evidence for Differentiation Theory

The remainder of this paper is an in-depth consideration of some of the evidence for Differentiation Theory. In Section 2 we describe and critique the triad task, re-analyze one of our previous experiments using this procedure (Milton et al., 2008, Experiment 5), and conduct three large-scale replications of two classic studies that also used this procedure (Smith & Kemler Nelson, 1984; Ward, 1983). In Section 3, we describe and critique the criterial-attribute procedure, and conduct three large-scale replications of two classic studies that used this procedure (Kemler Nelson, 1984; Smith & Shapiro, 1989). We then conclude with some general remarks about Combination and Differentiation Theory.

2. The Triad Task

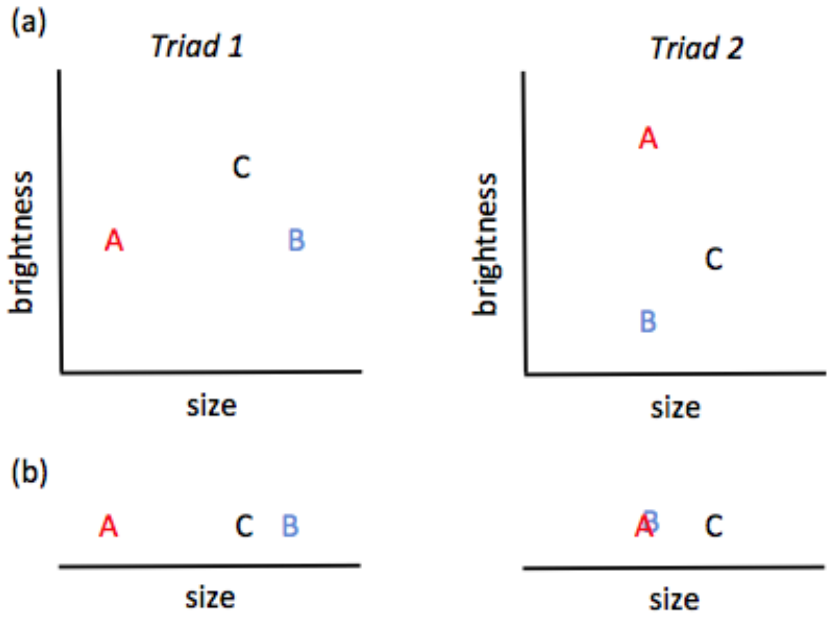

Figure 2: (a) Example of stimulus structure in the triad task. Typically, on half of trials A and B are identical on one dimension (e.g. brightness), and on the other half of trials A and B are identical on the other dimension (e.g. size). (b) Visualization of stimulus structure under the assumption brightness is ignored; in the rightmost panel, the overlapping of A and B is intended to indicate they occupy the same position on the size dimension. A, B and C are colored to make this overlap more visually obvious; the color conveys no other information.

2.1. Description and key results

For the purposes of the current article, the term *triad task* means any spontaneous classification procedure that involves the simultaneous presentation of three stimuli whose relationship to each other is as shown in Figure 2 and where the participant is required to state which two stimuli "go together best" (or which is the odd one out). Although the presented stimuli may differ from trial to trial, the relationship between them is constant; as illustrated in Figure 2, two of the stimuli (conventionally labelled A and B) are identical on one dimension, but very different on the other. At the same time, two stimuli (B and C) are identical on neither dimension, but similar on both, such that B and C are overall more similar to each other than are A and B.

In the standard terminology of the triad task, a participant who responds that B and C go together best is described as having made an *overall similarity* response. A participant who responds that A and B go together best is often described as having made a *dimensional* response. For reasons that will become clear later, we prefer the term *identity* response. The term refers to the fact that the two stimuli that are identical on one dimension have been classified together. The final possibility (responding that A and C go together best) is conventionally described as a *haphazard* response, as it is hard to see why participants would favor this response under any consistent classification strategy.

The triad task is discussed in Garner (1976, p.106) but, in terms of its use in the evaluation of Differentiation Theory, the seminal results (at least in the study of adults, which is the focus here) are those of Ward (1983), and Smith and Kemler Nelson (1984). Ward (1983) demonstrated that putting participants under time pressure increased the proportion of overall similarity (BC) responses they made, and reduced the proportion of identity (AB) responses. The interpretation of this result within Differentiation Theory is that time pressure results in participants making a decision on the basis of the undifferentiated whole that is the input to the classification process. This undifferentiated whole gives the participant rapid access to the overall similarity relationships between stimuli and hence results in a BC classification, because B and C are most similar overall. In the absence of time pressure, participants are able to differentiate the stimuli into their constituent attributes of length and dot-density and they then choose to weight the dimension containing an identity sufficiently that it drives responding (for a discussion of why identity is favored see Smith, 1989). Thus, as time pressure reduces, the proportion of identity (AB) responses increases and the

proportion of overall similarity (BC) responses falls.

Smith and Kemler Nelson (1984) substantially expanded upon Ward's initial demonstration. Using a size-brightness stimulus set, they replicated Ward's effect of time pressure on the triad task and, in addition, demonstrated that overall similarity (BC) responding could be increased by the introduction of a concurrent load (they also reported an instructional manipulation). The time pressure results of Ward (1983) and Smith and Kemler Nelson (1984) were recently replicated and extended by Milton et al. (2008, Experiment 5). We discuss the last of these studies in more detail in Section 2.6.

2.2. Single-dimension responding

In addition to overall similarity responding and identity responding, there is another simple strategy that participants might employ in the triad task—responding on the basis of a single dimension. This might take the form of, for example, consistently ignoring size and responding entirely on the basis of brightness. It might alternatively take the form of arbitrarily picking one dimension for each decision; for example, on some trials ignoring size and using brightness, on others ignoring brightness and using size. Either of these strategies will result in the participant making 50% AB responses and 50% BC responses. This is because, as illustrated in Figure 2, the dimension on which the identity occurs varies unpredictably from trial to trial (although it occurs equally often on the size and brightness dimensions across trials). For a participant classifying on the basis of size, when the identity occurs on the size dimension, they will of course respond that A and B are most similar (they are identical on the size dimension). However, when the identity occurs on the brightness dimension, a participant classifying on the basis of size will respond that B and C are most similar (because A and B are not very similar on the size dimension, but B and C are).

The possible presence of a third, on the face of it quite reasonable, response strategy introduces severe ambiguity into the traditional analysis of data from the triad task. Specifically, any single BC response may result either from classifying on the basis of overall similarity, or from classifying on the basis of a single dimension. Further, any single AB response may result either from (a) classifying on the basis of the dimension containing an identical match, which varies unpredictably from trial to trial, and thus involves consideration of both stimulus dimensions in order to determine on which dimension the identity occurs, or (b) classifying on the basis of a single dimension, ignoring

the other. The problem becomes particularly acute when one notes that AB responding tends to dominate, and BC responding is rare, in the absence of time pressure. Thus, if the effect of time pressure is to increase the likelihood that participants engage in single-dimension responding (as is observed in the match-to-standards procedure, e.g. Milton et al., 2008), then this will manifest as an increase in BC responding.

In summary, although an increase in BC responding in the triad task under time pressure has been taken to indicate an increase in overall similarity responding, it could also indicate an increase in single-dimension responding. Differentiation Theory predicts an increase in overall similarity responding as a result of time pressure; Combination Theory predicts an increase in single-dimension responding as a result of time pressure. Thus, the known effects of time pressure in the triad task are consistent with both theories.

2.3. Response-set analysis

In some ways, the triad task is similar to the match-to-standards task. In both tasks participants are simultaneously presented with three stimuli and are asked to choose which two go together best. In both tasks, they do so in the absence of feedback. Yet, despite these similarities, the results of the two procedures seem basically opposite. All the factors that increase overall similarity responding in the triad task (e.g. time pressure, concurrent load), decrease overall similarity responding in the match-to-standards task, and vice versa.

Of course, there are also a number of differences between the two procedures. In the match-to-standards task, two of the three stimuli are the same on every trial (the standards), and classifying the standards together is not permitted. The stimuli employed in the two procedures also differ — for example, in the number of stimulus dimensions and in the similarity relationship between the three stimuli. When we have discussed the discrepancy in conclusions between the match-to-standards and triad procedures with colleagues and anonymous reviewers, the consensus seems to be that some difference in stimulus construction or presentation is the underlying cause. In the current article, we make the case that the results of the two procedures are actually consistent with one another, and that the apparent discrepancy comes from the ambiguous analysis method traditionally employed with the triad task procedure.

In the traditional analysis, each response made by a participant is considered in isolation, and

one simply reports the proportion of AB, BC and AC responses made. As discussed above, this form of analysis is not particularly revealing because any single AB or BC response could have been the product of at least two different classification strategies. A more satisfactory approach is to not consider each response in isolation but to consider the full set of responses made by an individual. This technique, described here as *response-set analysis*, is routinely used in the match-to-standards procedure, and has also proven fruitful in both developmental (e.g. Thompson, 1994) and comparative (e.g. Wills et al., 2009) investigations of classification. In the current studies we apply this analysis to some of the classic studies that inspired those investigations.

In the following analyses, we consider three principal response strategies - Overall Similarity, Unidimensional, and Identity. The operation of these three response strategies is illustrated below through reference to the two triads in Figure 2:

Identity (ID). Participants place together those stimuli that share an identical attribute. For Triad 1, the participant therefore considers stimulus C to be the odd-one-out, as noted in Table 1. For the same reason, C is also the odd-one-out in Triad 2.

Overall Similarity (OS). Participants place together those stimuli that are overall most similar. This participant considers A as the odd-one-out for both Triad 1 and Triad 2.

Unidimensional (UD). This is a combination of two distinct participant strategies. The first strategy is that participants classify on the basis of size, ignoring brightness. For Triad 1, A is the odd-one-out on this strategy, while for Triad 2, C is the odd-one-out. The second strategy is that participants classify on the basis of brightness, ignoring size. Here, the predictions are opposite. Note that therefore participants who switch unpredictably between size and brightness from trial to trial are not easily detectable in this analysis. However, it is possible to detect participants who consistently base their response on one of the two dimensions. UD(size) and UD(brightness) are both single-dimension strategies and hence are categorized together as a UD strategy in our analysis.

Inspection of Table 1 shows that the ID, OS and UD response strategies become distinguishable when the responses to more than one stimulus triad are considered in combination. In the current studies, we evaluate, for each participant, the proportion of their responses predicted by each of the four response strategies in Table 1. We also include a response-bias strategy that pre-

	Triad 1	Triad 2
ID	C	C
OS	A	A
UD(size)	A	C
UD(brightness)	C	A

Table 1: Predicted "odd-one-out" responses for the triads in Figure 2, under Identity (ID), Overall Similarity (OS) and Unidimensional (UD) models.

dicts participants respond by pressing the same key on each trial. The strategy that predicts the highest proportion of responses for the participant is selected for that participant. In the current experiments, response-strategy *consistency* and *margin* were both substantial. Consistency is the proportion of the participant's responses predicted by the selected response strategy; mean consistency was 74%, against a chance level of 33%. Margin is the difference in the proportion of responses predicted by the selected (best) response strategy and the next-best strategy; mean margin in these experiments was 23%. More detailed analyses of consistency and margin and presented in the Supplementary Materials.

2.4. Predictions

Response-set analysis has the potential to reveal certain effects of time pressure that would disconfirm Combination Theory, others that would disconfirm Differentiation Theory, and others that would be consistent with both accounts. We consider these in turn below:

Disconfirmation of Combination Theory. Combination Theory assumes that the participant begins with the attributes and combines these if there is sufficient time. Therefore, the UD strategy should be the least time consuming of the three strategies we consider, as the other two involve consideration of both stimulus dimensions (the ID strategy involves consideration of both dimensions because the dimension containing the identity varies unpredictably from trial to trial). Thus, if time pressure is found to *reduce* single-dimension responding (reduce the prevalence of UD responders), this would disconfirm Combination Theory.

Disconfirmation of Differentiation Theory. Symmetrically, if time pressure leads to an increase in single-dimension responding, this would disconfirm Differentiation Theory. With less time, the individual attributes of the stimulus should be less available and thus it should be harder to, for example, respond on the basis of size while ignoring brightness.

Results consistent with both theories. There are a number of possible effects of time pressure that would be consistent with both Combination and Differentiation Theory although, naturally, their explanations of such results differ somewhat. Of particular relevance to the data in the current article is the possibility that time pressure might lead to a reduction in the prevalence of ID classifiers, an increase in the prevalence of OS classifiers, and no change in the prevalence of UD classifiers. Differentiation Theory explains this result by assuming that time pressure reduces the time for differentiation of the "blob" into its constituent attributes, and that direct access to the attributes is required in order to differentially weight the dimension on which the identity occurs. Combination Theory explains the same result by postulating that while combining attributes takes time, combining them in an unweighted manner (which leads to OS classification) take less time than combining them in a weighted manner (which is required for ID classification). The key point is that an increase in the prevalence of OS classifiers as a result of time pressure does not support Differentiation Theory over Combination Theory, as long as it is accompanied by a decrease in ID classification, rather than by a decrease in UD classification.

2.5. General Method

2.5.1. Participants and apparatus

Participants were recruited from the student populations at the University of Exeter and Plymouth University; they participated for course credit or cash payment. The experiments were conducted using the E-prime package running on standard PCs with 19-in. monitors (17-in. in Experiment 1). Responses were collected via standard keyboards.

2.5.2. Stimuli

The abstract stimulus structure is shown in Figure 3; physical stimulus dimensions varied between experiments. Eight stimulus triads are possible within this structure: 1-3-7, 1-5-7, 2-4-8, 2-6-8, 1-2-3, 1-2-4, 5-7-8 and 6-7-8. There are six different ways in which three stimuli can be placed

in three spatial locations, thus each of the eight triads had six different instantiations, resulting in 48 physically different triads per experiment.

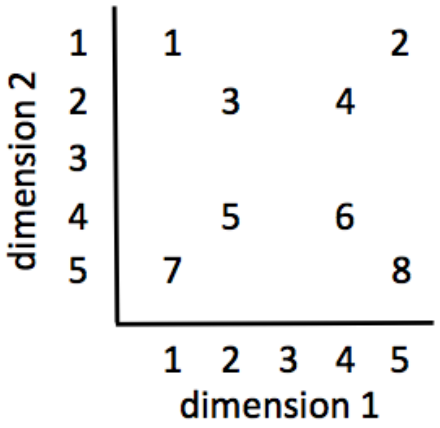

Figure 3: Abstract stimulus structure in Experiments 1-3.

2.5.3. Procedure

Each physically different triad was presented once (Experiment 1) or twice (Experiments 2-3), with order of presentation randomized for each participant. On each trial, the three stimuli of the presented triad were shown simultaneously, and participants pressed a key to indicate which stimulus they considered to be the "odd one out". An odd-one-out response was also employed by Ward (1983) and Smith and Kemler Nelson (1984). Time pressure was manipulated between subjects; the manner in which this was done varied between experiments, see subsequent sections for details.

In Experiment 2, Experiment 3A, and in the high time pressure condition of Experiment 3B, the classification phase described above was followed immediately by a standard self-report measure of impulsivity (Stanford, Mathias, Dougherty, Lake, Anderson & Patton, 2009). The results of the impulsivity questionnaires were not the main focus of the current investigation and are discussed in the Supplementary Materials.

2.6. Experiment 1

Our investigation commenced with a re-analysis of some of our own data on the effects of time pressure on classification in the triad task. These data were originally reported in Milton et al. (2008, Experiment 5) but, in that paper, the data were exclusively analyzed in the traditional

manner described in Section 2.2. Given the problems with this traditional method of analysis, our question was whether response-set analysis (Section 2.3) would support or overturn the conclusions of this study.

2.6.1. Method

Participants and Stimuli. 150 participants were tested. The stimuli were cartoon boats, 5.4 cm high by 7.7 cm wide, modeled on those used by Lamberts (1998). They varied on two dimensions, the length of the base of the hull and the length of the base of the sail. An example triad is shown in Figure 4A; the full set of eight physical stimuli instantiating the abstract structure described in Figure 3 are shown in Figure 5 of Milton et al. (2008).

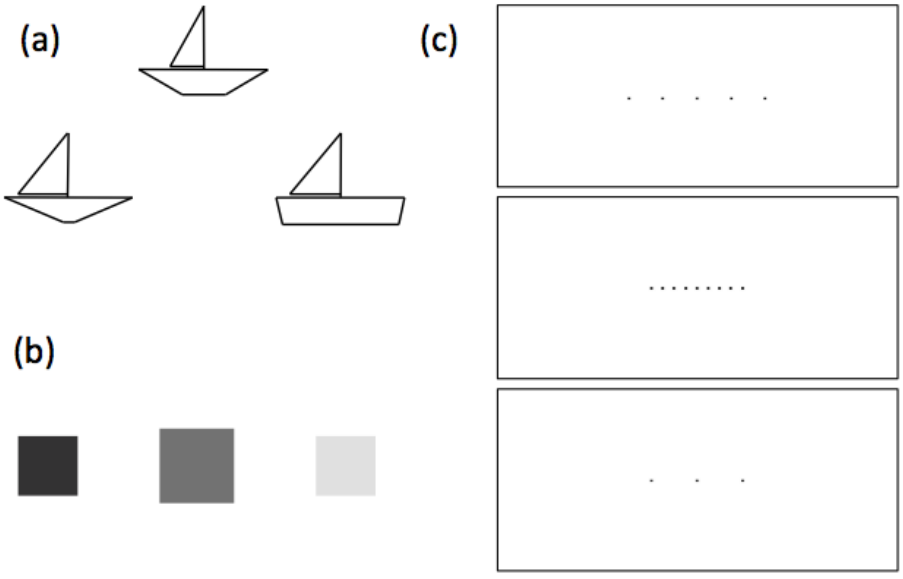

Figure 4: Example stimulus triads for (A) Experiment 1, (B) Experiment 2, (C) Experiments 3A and 3B. Examples are illustrative and the three panels are not drawn to the same scale.

Procedure. Participants were randomly assigned to one of five conditions, which were identical except for stimulus presentation time. The stimulus presentation times used were 640 ms, 1024 ms, 2048 ms, 3072 ms and 7500 ms. At the beginning of each trial a fixation cross was presented for 500 ms, followed by a blank screen for 500 ms. After this, the stimulus triad was presented for the appropriate duration and then immediately replaced by a mid-gray mask, along with the message "Please respond now". Participants used the left, right, and up, cursor keys to indicate which

stimulus they considered to be the odd one out. For example, if they thought the top stimulus in the triad was the odd one out, they pressed the up cursor key. The screen then went blank for 1500 ms before the next trial began.

Data archiving. The trial-level raw data is archived at `www.willslab.co.uk/exe1/` with md5 checksum `e3e824be2ff864c9cafbe55ec2646f3d`.[1]

2.6.2. Results and Discussion

Response-set analysis requires raw data down to the level of individual trials — a finer grain of raw data than is required for the traditional analysis (for which only proportions of AB, BC and AC responses for each participant are required). This finer grain of data had been retained in our archives for all but five participants; all subsequent analyses were performed on the 145 participants for which we had complete data. In the case of the response-set analysis, a further seven participants were excluded because they were best fitted by a response bias model (pressing the same key on each trial).

	Traditional			Response-set		
Condition	BC ("OS")	AB ("ID")	AC ("Hap.")	UD	OS	ID
640 ms	.43	.33	.24	.74	.22	.04
1024 ms	.49	.32	.19	.76	.24	.00
2048 ms	.44	.43	.13	.76	.10	.14
3072 ms	.35	.52	.13	.44	.20	.36
7500 ms	.38	.53	.09	.28	.36	.36

Table 2: Traditional and response-set analyses of Experiment 1. Traditional: proportion of BC, AB and AC responses, typically considered to represent Overall Similarity, Identity, and Haphazard responding, respectively. Response-set: Proportion of participants best fit by a Unidimensional (UD), Overall Similarity (OS), and Identity (ID) response model.

[1] Publication of an MD5 checksum allows the reader to independently confirm that the raw data in the archive is unchanged.

Table 2 shows the results of both the traditional and the response-set analyses of these data. Under the traditional analysis, the prevalence of "overall similarity" (BC) responding increases as presentation time decreases from 7500 ms to 1024 ms. The prevalence of BC responding is significantly higher at 1024 ms than it is at 3072 ms, $t(54) = 3.20, p < .01$, or at 7500 ms, $t(57) = 2.02, p < .05$. The prevalence of BC responding remains approximately stable between 1024ms and 2048 ms, $t(58) = 1.49, p = .14$, and between 3072 ms and 7500 ms, $t(53) < 1$. In summary, a comparison of the 1024 ms condition with either the 3072 ms condition or the 7500 ms condition indicates that time pressure increases the prevalence of "overall similarity" (BC) responding. Under the traditional analysis, this result provides support for Differentiation Theory.

However, response-set analysis (also presented in Table 2) reveals that the proportion of participants responding unidimensionally increased as presentation time decreased. The proportion of unidimensional responders is higher at 1024 ms than it is at 3072 ms, $\chi^2(1) = 5.74, p = .02$, or at 7500 ms, $\chi^2(1) = 12.78, p < .0005$. These results are predicted by Combination Theory but are inconsistent with Differentiation Theory. The increasing proportion of Unidimensional responders as time pressure increases explains the apparent increase in "overall similarity" (BC) classification under the traditional analysis—unidimensional responders emit 50% BC responses.

Inspection of Table 2 indicates a slight increase in true Overall Similarity responders at 1024 ms relative to 3072 ms, but this difference is not significant, $\chi^2(1) < 1$ and, in any case, the direction of the effect is consistent with both Combination and Differentiation theory (although for different reasons, see Section 2.4). The difference in proportion of Overall Similarity Responders between 1024 ms and 2048 ms is also not significant, $\chi^2(1) = 1.93, p = .16$. Contrary to the predictions of Differentiation theory, the proportion of Overall Similarity responders is higher at 7500 ms than at 1024 ms, although this trend is also not significant, $\chi^2(1) < 1$ (it is also not significant for the 2048 ms versus 3072 ms comparison, $\chi^2(1) < 1$). In summary, the response-set analysis reveals that the primary effect of time pressure in this experiment was to increase the proportion of participants responding on the basis of a single dimension. This result is predicted by Combination Theory but inconsistent with Differentiation Theory. In light of this radical revision to our previous conclusions about these data (Milton et al., 2008), we decided to investigate response-set analyses of some of the classic results within the triad task procedure.

2.7. Experiment 2

Smith and Kemler Nelson (1984, Experiment 3) reported that time pressure increased the prevalence of overall similarity classification in a triad task that employed a size-brightness stimulus set. Correspondence with the first author (J. D. Smith, personal communication, 3 Dec 2012) established that the raw data was not available, and that direct replication was not possible because there was no record of the specific stimulus triads employed. We thus ran a study using a similar range of size and brightness values to the original study, and constructed triads within this stimulus space using the abstract structure shown in Figure 3.

The manipulation of time pressure employed by Smith and Kemler Nelson (1984) was somewhat idiosyncratic, involving collapsing multiple, slightly different, small-sample conditions into a single condition. There was also an ineffective within-subjects manipulation of time pressure, and a failure to control the number of stimuli classified in speeded and unspeeded conditions (unspeeded participants classified more triads). This last aspect is particularly problematic, as Ward (1983) had demonstrated that AB responding can sometimes increase with increased experience of the task, and thus the observed difference between conditions could have arisen from this difference in experience rather than a difference in time pressure.

In the current study, we employed a simpler manipulation of time pressure, similar to that employed in Experiment 1. In our high time pressure condition, each triad was presented for one second. Experiment 1 had indicated that BC responding was maximal around 1 second presentation time, at least for the stimuli employed in that study. Also, while not directly comparable due to procedural differences, both our one-second presentation time, and Smith and Kemler Nelson's 32 decisions in 34 seconds procedure, put participants under substantial time pressure. In our low time pressure condition, each triad was presented for five seconds. This provided ample time to make each decision, and our previous work suggests that setting a fixed, long, duration has a similar effect to making the procedure self-paced (Milton et al., 2008).

2.7.1. Method

Participants and stimuli. 80 participants were tested. The stimuli were gray squares that varied in length of side (1.3, 1.5, 2.1 and 2.5 cm) and brightness (19%, 45%, 84% and 94%, where 0% and 100% are the minimum and maximum displayable brightnesses, respectively). Size values were taken from Smith and Kemler Nelson (1984) and the monitor was placed at a distance that

approximated the distance from eye to stimulus in a paper-based task. Brightness values were matched by eye by the first author to the standard Color Aid grayscale papers reported by Smith and Kemler Nelson(1984). The three stimuli making up a triad were presented in a horizontal line; an example triad is shown in Figure 4B.

Procedure. Participants were randomly assigned to one of two conditions, which were identical except for stimulus presentation time. The stimulus presentation times used were 1000 ms and 5000 ms. At the beginning of each trial, the screen displayed the message "Ready?" and the participant pressed a key to continue. After this, the stimulus triad was presented for the appropriate duration and then immediately replaced by the message "Odd one out?". The participant pressed the number key 1, 2 or 3 to indicate the left, middle, or right stimulus, respectively. The next trial began immediately upon detection of a response.

Data archiving. The trial-level raw data is archived at `www.willslab.co.uk/plym1/` with md5 checksum `06f1a097ad7635b2e1026e253d3969ab`.

	Traditional			Response-set		
Condition	BC	AB	AC	UD	OS	ID
1000 ms	.34	.55	.11	.38	.13	.49
5000 ms	.12	.85	.03	.10	.00	.90

Table 3: Traditional and response-set analyses of Experiment 2. Traditional: proportion of BC, AB and AC responses. Response-set: Proportion of participants best fit by a Unidimensional (UD), Overall Similarity (OS), and Identity (ID) response model.

2.7.2. Results and Discussion

Table 3 shows the results of both the traditional and response-set analyses. Under the traditional analysis, the prevalence of "overall similarity" (BC) responding is higher in the high time pressure (1000 ms) condition than in the low time pressure (5000 ms) condition, $t(78) = 5.80, p < .0001$, thus supporting the main conclusion of Smith and Kemler Nelson (1984).

Turning to the response-set analyses, two participants were best fit by a response bias model (i.e. always pressing the same key) and were excluded from further analysis. The critical result, shown in

Table 3, is that the proportion of participants responding unidimensionally increased as presentation time decreased, $\chi^2(1) = 8.64, p < .005$. This result is predicted by Combination Theory but is inconsistent with Differentiation Theory. The proportion of Overall Similarity responders was also higher in the 1000 ms condition than the 5000 ms condition, and this difference was significant, $\chi^2(1) = 5.92, p = .02$. This result is consistent with both Combination Theory and Differentiation Theory, although the explanation under these two theories is different (see Section 2.4).

2.8. Experiment 3A

Ward (1983, Experiment 2) reported that time pressure increased the prevalence of overall similarity classification in a triad task. Where Smith and Kemler Nelson (1984) used size-brightness stimuli, Ward employed dotted-line stimuli that varied in line length and dot density. Given the results of Experiments 1 and 2, it is tempting to conclude that Ward's result was also caused by an increase in the proportion of unidimensional responders under time pressure. However, an alternative possibility is that the effects of time pressure on the triad task are dependent on the specific combination of stimuli and time pressures employed.

Correspondence with Ward indicated that neither the raw data nor a list of the specific triads tested were available (T.B. Ward, personal communication, 8 Nov 2012). In Experiment 3A, we ran a large-scale conceptual replication of Ward's study, using a similar range of lengths and dot densities as the original study, and constructing triads within this stimulus space using the abstract structure shown in Figure 3.

Ward's time pressure manipulation was to permit a maximum of two seconds per decision in the high time pressure condition, and insist on a minimum of five seconds per decision in the low time pressure condition. In Experiment 3A, as in Experiments 1 and 2, we manipulated time pressure by changing the duration of stimulus presentation. We chose stimulus presentation times of 2000 ms and 5000 ms in order to approximate the levels of time pressure in Ward's study.

The relatively small number of dots making up some of the stimuli raises the possibility of a further simple strategy that participants might employ — counting the number of dots in each stimulus and classifying on that basis. Ward (1983) had also considered this possibility and concluded that there was little evidence his participants were using such a strategy. Nevertheless, we included this Number strategy in our response-set analysis of both Experiment 3A and Experiment 3B.

2.8.1. Method

Participants and stimuli. 107 participants were tested. The stimuli were dotted horizontal lines that varied in length (2, 3, 6 and 8 cm) and inter-dot distance (1, .75, .375, and .25 cm). These values were selected from the range of values employed by Ward (1983), and the monitor was placed at a distance that approximated the distance from eye to stimulus in a paper-based task. The three stimuli making up a triad were presented inside rectangles of fixed size, placed one above the other, as illustrated in Figure 4C. This presentation format was chosen to emulate the presentation of stimuli on index cards (the procedure employed by Ward).

Procedure. The procedure was identical to Experiment 2, except that the high time pressure condition employed a 2000 ms presentation time (rather than 1000 ms). Responses were collected using the O, K, and M keys, reflecting the vertical arrangement of the stimulus triad.

Data archiving. The trial-level raw data is archived at `www.willslab.co.uk/plym2/` with md5 checksum `6a2b75178ae4ee334f0c63e4407bb357`

2.8.2. Results and Discussion

Table 4 shows the results of both the traditional and response-set analyses. Under the traditional analysis, the prevalence of "overall similarity" (BC) responding is higher in the high time pressure (2000 ms) condition than in the low time pressure (5000 ms) condition, $t(105) = 2.03, p = .045$, thus supporting the main conclusion of Ward (1983).

Turning to the response-set analysis, three participants were best fit by the Number model; these participants were excluded from further analysis. Inspection of Table 4 shows that the main effect

	Traditional			Response-set		
Condition	BC	AB	AC	UD	OS	ID
2000 ms	.66	.20	.14	.29	.67	.04
5000 ms	.58	.32	.10	.40	.42	.18

Table 4: Traditional and response-set analyses of Experiment 3A. Traditional: proportion of BC, AB and AC responses. Response-set: Proportion of participants best fit by a Unidimensional (UD), Overall Similarity (OS), and Identity (ID) response model.

of time pressure in this study was to increase the proportion of Overall Similarity responders. This increase, which is significant, $\chi^2(1) = 6.38, p = .01$, is consistent with both Combination Theory and Differentiation Theory. Time pressure also seems to decrease the proportion of Unidimensional responders, although this effect, despite the considerable sample size of this study, does not approach significance, $\chi^2(1) = 1.23, p = .27$.

2.9. Experiment 3B

Experiment 3B served two purposes. First, it addressed the possibility that the central result of Experiment 3A (increased proportion of overall similarity responders under time pressure) was in some way contingent on the particular manner in which time pressure had been manipulated. Ward (1983) manipulated overall decision time per triad, while Experiment 3A manipulated stimulus presentation time, leaving the decision itself self-paced. In Experiment 3B, we developed a close analog of Ward's time pressure manipulation within our computerized task in order to address this possibility. The second purpose of Experiment 3B was to examine whether the non-significant trend for a decreased proportion of Unidimensional responders under time pressure observed in Experiment 3A could be found significantly in a study that was a closer approximation to Ward's original procedure. Evidence of a significant reduction in the proportion of Unidimensional responders under time pressure would be inconsistent with Combination Theory.

2.9.1. Method

Participants and stimuli. 79 participants were tested. The stimuli were identical to those in Experiment 3A.

Procedure. The procedure was identical to that in Experiment 3A, except for the manner in which time pressure was manipulated. Each trial commenced with the message "Ready?" and the participant pressed a key to continue. In the < 2000 ms condition, the stimuli were presented simultaneously with the message "odd one out?". If participants did not respond within 2000 ms of stimulus onset, the stimuli were removed from the screen and the next trial began immediately. In the > 5000 ms condition, the stimuli were presented simultaneously with the message "study carefully" for 5000 ms. Responses during this period were ignored. After 5000 ms, the "study carefully" message was replaced with the message "when you are ready ... odd one out?". The stimuli remained on the screen until the participant made a response.

Data archiving. The trial-level raw data is archived at `www.willslab.co.uk/plym3/` with md5 checksum `4eb35fc660c8a6569d77988feee8a75e`

2.9.2. Results and Discussion

Table 5 shows the results of both the traditional and response-set analyses. Under the traditional analysis, the prevalence of BC responding is higher in the high time pressure condition (< 2000 ms) than in the low time pressure condition (> 5000 ms), although this difference is not significant, $t(77) = 1.06, p = .29$. The lack of statistical significance in the traditional analysis is of relatively little consequence, given the interpretive ambiguity inherent in this method of analysis.

Turning to the response-set analysis, fourteen participants were best fit by the Number model (seven participants in each condition). These participants, along with one participant best fit by a response bias model, were excluded from further analysis. Inspection of Table 5 indicates that the main effect of time pressure in the current experiment was to increase the proportion of Overall Similarity responders. This increase was significant, $\chi^2(1) = 7.73, p = .005$, and is consistent with both Combination Theory and Differentiation Theory. There was also a trend for time pressure to reduce the proportion of Unidimensional responders. However, as in Experiment 3A, this difference was not significant, $\chi^2(1) = 1.90, p = .17$. Combining Experiments 3A and 3B into a single analysis, the effect still fell short of significance, $\chi^2(1) = 2.48, p = .12$. Bayesian analysis indicates that the null hypothesis is more than three times as likely as the experimental hypothesis for these combined data, $BF = .30$, estimated with the `bcct` function of the package `conting` (Overstall, 2014; Overstall & King, 2014) in the `R` environment (R Core Team, 2014). $BF < .33$ is typically considered to be substantial evidence for the null (Jeffreys, 1961).

	Traditional			Response-set		
Condition	BC	AB	AC	UD	OS	ID
< 2000 ms	.63	.13	.24	.09	.88	.03
> 5000 ms	.58	.32	.10	.21	.56	.22

Table 5: Traditional and response-set analyses of Experiment 3B. Traditional: proportion of BC, AB and AC responses. Response-set: Proportion of participants best fit by a Unidimensional (UD), Overall Similarity (OS), and Identity (ID) response model.

2.10. Interim Summary

Across four experiments employing the triad task procedure, we found some results predicted by Combination Theory but inconsistent with Differentiation Theory, and other results consistent with both accounts. We found no effect that supported Differentiation Theory over Combination Theory. The results of these studies thus favor Combination Theory over Differentiation Theory. Our conclusion is opposite to that drawn by Smith and Kemler Nelson (1984) and by Ward (1983). The difference between our conclusions and theirs seems unlikely to be due to substantive differences in procedure or population because, when we follow their method of analysis, we obtain the same ordinal effects that they did. In the next section, we turn our attention to another set of studies taken to support Differentiation Theory over Combination Theory—those employing the criterial-attribute procedure.

3. The Criterial Attribute Procedure

3.1. Description and key results

The term *criterial attribute procedure*, as defined here, refers to any supervised category learning procedure that employs the abstract stimulus structure shown in Table 6. Category learning procedures involve the, usually sequential, presentation of unfamiliar stimuli; supervised category learning procedures are those where information about the category membership of each stimulus is available on each trial. The critical aspect of the criterial attribute procedure is that, during training, participants can achieve perfect performance in two distinct ways. First, they can discover that one stimulus dimension (dimension 1 in the case of Table 6) is perfectly predictive of category membership during training, and respond on the basis of that dimension, ignoring the others. This is described as *criterial attribute* responding. An alternative strategy participants could employ to achieve perfect responding is to classify on the basis of overall similarity. For example, the stimulus `1101` is a member of category A because it is more similar to (shares more features with) members of category A than category B. This is described as *overall similarity* responding. These two strategies can be distinguished during the test phase through what are described as *critical test stimuli*. Criterial-attribute responding and overall similarity responding lead to opposite responses to these stimuli. For example, consider stimulus `1000` in Table 6. Criterial-attribute responding places this stimulus into category A, because a value of zero on dimension 1 perfectly predicted

	Stimulus				Response model				
	D1	*D2*	*D3*	*D4*	*CA*	*N2*	*N3*	*N4*	*OS*
Train A	1	1	1	1	A	A	A	A	A
	1	1	1	0	A	A	A	B	A
	1	1	0	1	A	A	B	A	A
	1	0	1	1	A	B	A	A	A
Train B	0	0	0	0	B	B	B	B	B
	0	0	0	1	B	B	B	A	B
	0	0	1	0	B	B	A	B	B
	0	1	0	0	B	A	B	B	B
"Critical" Test	0	1	1	1	B	A	A	A	A
	1	0	0	0	A	B	B	B	B
Other Test	0	0	1	1	B	B	A	A	?
	1	1	0	0	A	A	B	B	?
	0	1	0	1	B	A	B	A	?
	1	0	1	0	A	B	A	B	?
	0	1	1	0	B	A	A	B	?
	1	0	0	1	A	B	B	A	?

Table 6: Abstract stimulus structure for criterial-attribute procedurte employed in Experiments 4A, 4B, and 5. Each stimulus comprises four binary dimensions D1–D4, see Figure 5. *Train A* are the members of category A presented during the training phase, similarly for *Train B*. *"Critical" test* are the stimuli presented during test that are traditionally considered to be theoretically critical. *Other test* are the other stimuli presented during the test phase. The column *CA* indicates how each stimulus would be classified under a strategy of responding to the Criterial Attribute (the dimension that perfectly predicts category membership during training — dimension 1 in this example). Columns *N2, N3, N4* give the same information under a non-criterial single-attribute strategy of responding to the second, third and fourth dimension, respectively. Column *OS* gives the predictions of an Overall Similarity strategy. A indicates a Category A response is predicted, B indicates a Category B response. ? indicates that a random response is predicted.

category A during training. Overall similarity responding, however, places stimulus `1000` into category B, because `1000` shares more features with the members of category B than with the members of category A.

The criterial-attribute procedure was introduced by Kemler Nelson (1984). In the sections that follow, we re-examine two key results within this procedure. The first is that incidental training, relative to intentional training, increases the prevalence of overall similarity responding (Kemler Nelson, 1984, Experiment 1). Under intentional training, participants are explicitly told that a category structure exists, and that their job is to discover it. Participants assign a category label to each presented stimulus, and get immediate feedback on the accuracy of each response. Incidental training also provides trial-specific information about category membership, but the training phase involves some task other than categorization, and participants do not have their attention explicitly drawn to the availability of category information during training. In Kemler Nelson (1984, Experiment 1), intentional training involved learning to classify schematic faces as belonging to either policemen or doctors. Incidental training involved presenting the same faces but instead asking whether each face had been seen previously. Category membership information was available in the incidental condition because each face was presented atop a policeman's or doctor's uniform, but no classification responses were required, and participants' attention was not explicitly drawn to the uniforms until the test phase. Both intentionally-trained and incidentally-trained participants were then tested under intentional conditions. Kemler Nelson (1984) concluded that participants in the incidental training condition were more likely than participants in the intentional condition to respond to the critical test stimuli on the basis of overall similarity. This result is consistent with Differentiation Theory under the assumption that participants dedicate fewer cognitive resources to classification under incidental than intentional conditions. Fewer cognitive resources lead to an increased likelihood of overall similarity classification under Differentiation Theory because such resources are required to split the undifferentiated whole into its constituent attributes, which is in turn required to classify the stimulus on the basis of a single attribute.

The second key result we re-examine in the following sections is that concurrent load during intentional training increases the prevalence of overall-similarity-consistent responding to the critical test items in a subsequent full-attention test (Smith & Shapiro, 1989, Experiment 1). This result is consistent with Differentiation Theory, and inconsistent with Combination Theory, under the

assumption that concurrent load reduces the availability of cognitive resources for classification.

3.2. Non-criterial attribute responding

In addition to responding on the basis of the criterial attribute, or responding on the basis of overall similarity, there is a third simple strategy participants might employ—classifying on the basis of a single, non-criterial dimension. For example, consider a participant who responds on the basis of dimension 2 in Table 6. This participant can respond with 75% accuracy during training if they classify solely on the basis of this dimension. Although a less-than-perfect strategy, it is sufficient to pass the learning criterion in both Kemler Nelson (1984), and Smith and Shapiro (1989). Critically, the response to the critical test items under such a *non-criterial attribute* strategy is indistinguishable from the response under an overall similarity strategy. For example, a participant responding on the basis of dimension 2 in Figure 6 places stimulus `1000` into category B, just as a participant responding on the basis of overall similarity would. Hence, the criterial-attribute procedure has similar interpretative problems to the triad procedure, and for similar reasons. Specifically, responding to the critical test items on the basis of overall similarity is indistinguishable from responding on the basis of a single (non-criterial) dimension.

Kemler Nelson (1984) was aware of this problem, and took some steps to try to eliminate a non-criterial attribute explanation for her results. The final part of her test phase involved the presentation of all 16 possible stimuli, see Table 6. Examination of responses to this larger set of stimuli can distinguish overall similarity responding from non-criterial single-attribute responding. Kemler Nelson reported that only 1 of the 15 incidental participants were *perfectly* fit by a non-criterial single-attribute strategy. This was taken as evidence that non-criterial attribute responding did not explain her results, but the assumption that a model has to fit the data perfectly to be considered seems restrictive. From the data she reported (Kemler Nelson, 1984, p. 740), the mean consistency of a non-criterial attribute model in the incidental condition can be calculated as .94. The same figures allow us to conclude that the consistency of an overall similarity model on the critical test items alone could not have exceeded .84 in the incidental condition. Thus, there is some reason to suspect that non-criterial attribute responding may have provided a better explanation of participants' behavior in the incidental condition of this study than overall similarity responding.

Smith and Shapiro also considered the possibility of non-criterial attribute responding and, in response, reported their data in terms of the percentage of "true" or "strong" family resemblance

responders (Smith & Shapiro, 1989, p. 390). They defined strong family responders as those whose consistency with an overall similarity model on the critical test items alone exceeded .80, while their consistency with a single-attribute model across all items did not exceed .89. Note that responding on the critical test items does not differentiate between overall similarity and non-criterial attribute responding. Thus, any non-criterial attribute responder who was below .90 consistency across all test items would be considered a "true FR" responder on this metric (as long as they classify 80%+ of the critical test items in a way consistent with overall similarity / non-criterial attribute responding). In other words, it is possible that the proportion of "true FR" responders includes some participants who are moderately consistent in applying a non-criterial single-attribute strategy. If concurrent load acts to increase the proportion of participants who adopt a moderately consistent non-criterial attribute strategy, then the "true FR" proportion would increase under concurrent load, without any need to assume an increase in the proportion of people best fit by an overall similarity model.

The issue of non-criterial single-attribute responding under incidental training conditions has already been a topic of some debate between Ward and Kemler Nelson (Kemler Nelson, 1984; Ward & Scott, 1987; Ward, 1988). The details of this debate, which concerns a slightly different procedure to the one currently under consideration, are considered in the Supplementary Materials. The critical point is that, throughout this debate, both sides repeatedly asserted that non-criterial single-attribute responding could not explain the results of Kemler Nelson (1984, Experiment 1), the experiment on which the studies in the current paper are based. We believe this shared assertion may have been premature.

In summary, Kemler Nelson (1984), and Smith and Shapiro (1989), both considered and rejected the possibility that incidental training (or concurrent load) increases the proportion of participants best fit by a non-criterial attribute model. They favored the conclusion that incidental training (or concurrent load) instead increases the proportion of participants best fit by an overall similarity model. However, their rejection of the non-criterial attribute explanation was based on somewhat inconclusive analyses that employed some aspects of response-set analysis but stopped short of a full assessment.

In the sections that follow, we re-examine these two classic experiments using a full response-set assessment of the participants' trial-by-trial response to the test items. We did not seek the raw

data of the original studies, because the original studies had a very small number of test trials on the full stimulus set (in fact, the majority of the 16 possible test stimuli were presented just once per participant). Response-set analysis is likely to be more robust if performed over a larger number of test trials, as in the triad task. The original studies, particularly Kemler Nelson (1984), also had relatively small sample sizes. We thus ran large-scale replications of the original studies, with extended test phases in order to facilitate response-set analysis.

3.3. Response-set analysis

Response-set analysis of the criterial-attribute procedure is similar to response-set analysis of the triad task. Table 6 lists all 16 stimuli it is possible to generate from four binary stimulus dimensions; all these stimuli are presented in the test phase repeatedly and in random order. We consider three response strategies—Overall Similarity, Criterial Attribute, and Non-criterial Attribute. The operation of these three response strategies are discussed below.

Overall Similarity (OS). Participants place each test stimulus into the category for which it is overall most similar. For example, stimulus `1000` is placed into category B because it has three of the four features characteristic of category B and only one of the features characteristic of category A. Where a stimulus has an equal number of A-characteristic and B-characteristic features (e.g `1100`), the participant guesses (randomly selects A or B).

Criterial Attribute (CA). During training, participants identify the attribute that perfectly predicts category membership (dimension 1 in Table 6) and classify each test stimulus on the basis of that attribute alone. For example, stimulus `1000` is placed into category A because a value of 1 on dimension 1 perfectly predicted category A during training.

Non-criterial Attribute (NCA). This is a combination of three distinct strategies, one for each of the three stimulus dimensions that were characteristic rather than criterial during training. In Table 6 the non-criterial dimensions are 2, 3, and 4. During test, the participant responds on the basis of one of these non-criterial dimensions, ignoring the others. For example, a participant responding on the basis of dimension 2 places the stimulus `1000` into category B.

Inspection of Table 6 shows that OS, CA, and NCA response strategies become distinguishable when the responses to the full set of test stimuli are considered in combination (no column in Table

6 is identical to any other column). In the current studies we evaluate, for each participant, the proportion of their responses predicted by each of the five response strategies in Table 6. We also include a response-bias strategy that predicts participants respond by pressing the same key on each trial. The strategy that predicts the highest proportion of responses for the participant is selected for that participant. In the current experiments, the selected response strategy predicts a substantial proportion of the participant's responses (mean consistency: 86%, against a chance level of 50%), and the margin between the selected strategy and its nearest competitor is also substantial (mean margin: 13%).

The fact that the Overall Similarity model predicts a random response for some stimuli (e.g. `1100`) introduces a slight complication. Our default approach in the current studies was to note that these random responses will, on average, correspond to the participant's response on 50% of occasions and to score the model's responses accordingly. In other words, the Overall Similarity model is considered to have correctly predicted half of the responses to which it emitted a random response. This approach has the advantage of simplicity, but might be considered to disadvantage the Overall Similarity model. This is because the OS model can never predict all the participant's responses, even if it is the correct model (it has a maximal consistency of 81.25%).

There are at least three ways to mitigate the disadvantage of the Overall Similarity model. First, one can exclude the OS-ambiguous items (e.g. `1010`) from the response-set analysis. This approach is examined in Experiment 4A and Experiment 5. Second, one can remove OS-ambiguous stimuli from the experiment entirely, never presenting them to participants. This approach is examined in Experiment 4B. Third, one can move beyond response-set analysis to yet more sophisticated analyses. This approach is taken in the Supplementary Materials (Section 2), where we use logistic regression to examine the possibilities that (a) some people use more than one, but less than four, stimulus dimensions, and (b) people using more than one dimension may not weight those dimensions equally. This more complex analysis leads to the same conclusions as the simpler analysis reported in the main paper.

3.4. Predictions

Response-set analysis has the potential to reveal certain effects of incidental training, and concurrent load, that would disconfirm Combination Theory, others that would disconfirm Differentiation Theory, and others that would be consistent with both accounts. We consider these in

turn below.

Results consistent with both theories. In considering the predictions of either theory, it is important to remember that the presented stimuli have four experimenter-defined attributes, three of which predict category membership of the training items with 75% accuracy, while the fourth is a perfect predictor. Under both Combination Theory and Differentiation Theory, discovering the perfect predictor among three .75 predictors is a task that seems likely to require some cognitive resources. It therefore seems reasonable to assume under either theory that reliably discovering the criterial attribute is more effortful than responding on the basis of an arbitrarily chosen single dimension (which will be non-criterial with probability .75). If the effect of incidental training (or concurrent load) was to decrease Criterial Attribute responding while increasing Non-criterial Attribute responding, this would be consistent with both theories.

Disconfirmation of Differentiation Theory. Differentiation Theory assumes that the participant begins with a undifferentiated "whole" and separates this into its constituent attributes if there are sufficient resources to do so. Thus, Overall Similarity responding is less effortful than both criterial-attribute responding and non-criterial single-attribute responding. Differentiation Theory predicts that concurrent load and incidental training manipulations will increase Overall Similarity responding, and is thus disconfirmed if these manipulations decrease the prevalence of Overall Similarity responding.

Disconfirmation of Combination Theory. Combination Theory assumes that the participant begins with the attributes and combines these if there are sufficient resources to do so. Responding on the basis of Overall Similarity is thus more effortful than responding on the basis of a single dimension. It also seems reasonable to assume that reliably discovering the criterial attribute is more effortful than responding on the basis of an arbitrarily chosen single dimension (which will be non-criterial with probability .75). Therefore, Non-criterial Attribute responding is, overall, the least effortful of the strategies. Combination Theory is thus disconfirmed if incidental training or concurrent load decrease the prevalence of Non-criterial Attribute responding.

3.5. Experiment 4A

Experiment 4A was a replication of the incidental and intentional training procedures of Kemler Nelson (1984, Experiment 1), followed immediately by an extended test phase. As in Kem-

ler Nelson (1984), the experiment involved the classification of cartoon faces as belonging to either a doctor or a policeman.

3.5.1. Method

Participants and Apparatus. 106 undergraduate psychology students at the University of Plymouth participated for course credit. The experiments were conducted using the E-prime package running on standard PCs with 19-in. monitors. Participants sat approximately 50 cm from the screen.

Figure 5: Example stimuli for Experiments 4A and 4B.

Stimuli. The face stimuli were created by tracing Figure 3 of Kemler Nelson (1984). These tracings were then digitally scanned and the full stimulus set re-created through digital cut-and-paste procedures. Examples of the resulting images are shown in Figure 5. The stimuli varied on four dimensions; hair (straight or curly), ears (large or small), nose (large or small) and mustache (large or small). Each face was approximately 8 cm wide and 7 cm tall. The experiment also employed cartoon drawings of the top half of a policeman's uniform and the top half of a doctor's uniform; faces were placed atop these uniforms in order to indicate category membership. Both uniforms had a height of 4.5cm; the doctor's had a width of 5 cm and the policeman's had a width of 9 cm.

Procedure. Participants were arbitrarily assigned to either the intentional training condition or the incidental training condition. In both cases, training was followed immediately by an intentional test phase. The allocation of the criterial-attribute dimension (D1 in Table 6) to the physical dimensions (hair, ears, nose, mustache) was randomized across participants. In addition, each criterial-attribute subcondition was associated, across participants, with two different, randomly-selected, allocations of the feature levels of each dimension to the categories (doctor or policeman). For example, for one participant a large nose might be characteristic of doctors, while for another participant it might be characteristic of policemen.

Intentional training. Participants were told that their task was to categorize each presented face as either a doctor or a policeman, and that they should try to figure out how to tell which it was. Each trial started with the presentation of the two uniforms, to the bottom left and bottom right of the screen. The side of the screen on which each uniform appeared was counterbalanced across participants. After 1000 ms, the face stimulus appeared in the center of the top of the screen above the uniforms. After a further 2000 ms, "C - Policeman" and "M - Doctor" appeared under the uniforms, indicating the response keys to be used. Once the participant had responded, feedback was presented for 1500 ms. Feedback took the form of the face transposing onto the correct uniform and either the word "correct" or "incorrect" appearing at the top of the screen. The training phase contained three blocks, each block comprising one presentation of each of the eight training stimuli listed in Table 6, in random order.

Incidental training. Participants were told that their task was to decide whether or not they had previously seen each presented face, and rate their confidence in each of their decisions. No feedback was given. Each trial began with the presentation of the two uniforms, to the bottom left and bottom right of the screen (counterbalanced across participants). After 1000 ms, the face stimulus appeared atop the correct uniform. After a further 2000 ms, the question "Seen before? (Y/N)" appeared in the center of the bottom of the screen. After a response was made, the question "Confidence? 1 - 2 - 3" appeared at the center of the bottom of the screen (1 being the least confident, 3 being the most confident). The training phase contained three blocks, each block comprising one presentation of each of the eight training stimuli listed in Table 6, in random order. This training procedure, which is similar to that employed by Kemler Nelson (1984), may incidentally draw some attention to category membership of the faces, but no explicit mention of category membership was made by the experimenter, and the participants' task did not involve or require categorization.

Test phase. Participants in the intentional condition were informed that they would be performing the same task, but without feedback. Participants in the incidental condition were told that they may have noticed that in the training phase some of the faces belonged to doctors and some to policemen. They were told that they would now see more faces and that they should categorize them as either doctors or policemen. The procedure and timings of each test trial were identical to a trial in the training phase of the intentional condition, except that feedback was not given. The

test phase contained eight blocks, with each block comprising one presentation, in random order, of all sixteen stimuli it is possible to create from these binary four-dimensional stimuli.

Data archiving. The trial-level raw data is archived at `www.willslab.co.uk/plym5` with md5 checksum `32b3dcf8b578460c27751b625184f309`

3.5.2. Results and Discussion

Kemler Nelson (1984) excluded all participants who responded correctly on less than two-thirds of the training stimuli presented in the test phase. We applied the same two-thirds correct criterion to our participants, using the training stimuli from the first three test blocks of our test phase in order to approximately match the number of test-phase training stimuli over which performance was calculated. One might otherwise argue that our criterion was more stringent than Kemler Nelson's on the basis that some forgetting might occur during our longer test phase.

One idiosyncratic aspect of Kemler Nelson's design was that during testing, but not during training, the prototypes (`0000` and `1111`) were presented twice as often as the other training stimuli. Rather than replicating this aspect of her design, we opted instead to calculate proportion correct at test using a weighted average, assigning a weight of 2 to the prototypes and a weight of 1 to the other training stimuli. We believe that the way we have defined the learning criterion in the current study is as close to the criterion employed by Kemler Nelson as is possible given the slight differences in procedure between the two studies. Nevertheless, using an unweighted average leads to the same conclusions, as does defining the criterion over the whole test phase (rather than the first three blocks).

After applying the learning criterion, there were 29 participants remaining in the intentional condition and 30 participants remaining in the incidental condition. In order to achieve these approximately equal group sizes, more participants had to be run in the incidental condition than the intentional condition (35 intentional, 71 incidental). This was also the case in Kemler Nelson (1984), although the difference in pass rates was smaller in their study than in the current experiment. One further participant was excluded from the response-set analysis on the grounds that their best-fitting model was a key bias (pressing the same key on every trial).

The top two rows of Table 7 shows the results of both the traditional and the response-set analyses of these data. Under the traditional analysis, the proportion of "overall similarity" responses

		Traditional			Response-set		
	Condition	OS	CA		OS	CA	NCA
Passed	Intentional	.32	.68		.07	.68	.25
	Incidental	.52	.48		.03	.44	.53
ALL	Intentional	.34	.66	*All-items*	.06	.66	.28
				10-items	.12	.66	.22
	Incidental	.52	.48	*All-items*	.01	.34	.65
				10-items	.11	.27	.62

Table 7: Traditional and response-set analyses of Experiment 4A, for participants passing the learning criterion (Passed), and for all participants (ALL). Traditional: Proportion of Overall Similarity (OS) and Criterial Attribute (CA) responses to the critical test stimuli. Response-set: Proportion of participants best fit by an Overall Similarity (OS), Criterial Attribute (CA), and Non-criterial Attribute (NCA) response model. All-participants response-set analysis is reported for all test items (All-items), and for the 10 test items for which the OS response model makes a clear prediction (10-items).

to the critical test stimuli was higher in the incidental condition than in the intentional condition, $t(57) = 2.14, p = .04$. This replicates Kemler Nelson's central result and appears to support Differentiation Theory over Combination Theory. However, the response-set analysis leads to a different conclusion.

Response-set analysis, unlike the traditional analysis, can distinguish Overall Similarity responding from responding on the basis of a single, non-criterial, attribute. Table 7 reveals that Overall Similarity responders are rare under both intentional and incidental training conditions. There is a trend for the proportion of Overall Similarity responders to reduce under incidental training, relative to intentional training, but this trend is non-significant, $\chi^2(1) < 1$. The main result is a higher proportion of Non-criterial Attribute responders under incidental training than under intentional training, $\chi^2(1) = 4.86, p = .03$. The complementary difference in the proportion of Criterial Attribute responders is marginally significant, $\chi^2(1) = 3.52, p = .06$.

The conclusions of the response-set analysis are unchanged if one considers all participants, rather than just those who passed the learning criterion (see the bottom part of Table 7, in the rows labelled "ALL"). Specifically, there is a significantly higher proportion of Non-criterial Attribute re-

sponders under incidental conditions than intentional conditions, $\chi^2(1) = 11.45, p < .001$, and there is, in complementary fashion, a significantly higher proportion of Criterial Attribute responders under intentional conditions than under incidental conditions, $\chi^2(1) = 8.78, p = .003$. There is no significant effect of condition on the prevalence of Overall Similarity responders, $\chi^2(1) = 1.59, p = .26$. The effect in the traditional analysis remains significant, $t(104) = 2.51, p = .01$.

The conclusions of the response-set analysis are also unchanged if one considers only those stimuli for which the Overall Similarity model makes a clear prediction (see Section 3.3). The bottom part of Table 7, in the rows labelled "10-items", shows that the Overall Similarity response model provides the best fit for a slightly higher proportion of participants under this analysis, but Overall Similarity responders are still rare in both the intentional and incidental conditions. The trend for there to be a lower proportion of Overall Similarity responders under incidental conditions is not significant, $\chi^2(1) < 1$. It remains the case that the main result of the current experiment is a higher proportion of non-criterial attribute responders in the incidental condition than in the intentional condition, $\chi^2(1) = 12.95, p < .001$, and a complementary difference in Criterial Attribute responders, $\chi^2(1) = 12.27, p < .001$.

In summary, the effect of incidental training seems to be to make it harder for participants to identify the one stimulus attribute that perfectly predicts category membership (the criterial attribute), from among three other attributes that predict membership with 75% accuracy (the non-criterial attributes). However, this increased difficulty does not seem to increase the likelihood that participants will employ an overall similarity strategy. Rather, it simply increases the likelihood that the single attribute they use is not the criterial attribute.

3.6. Experiment 4B

In Experiment 4A, 38% of the test items were stimuli for which there is no clear answer from the perspective of unweighted Overall Similarity responding. This raises the possibility that participants were discouraged from applying an overall similarity strategy in the test phase because it gave no clear answer for a substantial minority of the test stimuli. In Experiment 4B, we addressed this possibility by removing the overall-similarity-ambiguous test items from the experiment entirely. Response-set analysis can still distinguish overall similarity, criterial attribute, and non-criterial attribute strategies on the basis of the participants' responses to the remaining 10 items. However, sensitivity of the analysis is somewhat reduced because the set of six OS-ambiguous stimuli provide

information that helps to distinguish Criterial Attribute from Non-criterial Attribute strategies (see Table 6). It is also notable that, in Experiment 4A, the pass rate for incidental training was much lower than the pass rate for intentional training. Differences in pass rate can complicate the interpretation of learning studies (Newell, Dunn & Kalish, 2010). In Experiment 4B, we sought to equate pass rates by extending the length of the training phase in the incidental condition.

3.6.1. Method

Participants, apparatus and stimuli. 75 undergraduate psychology students at the University of Plymouth participated for course credit, on apparatus identical to Experiment 4A. The stimuli were identical to Experiment 4A, except that the six overall-similarity-ambiguous stimuli (e.g. `1010`) were not presented.

Procedure. The intentional training phase was identical to Experiment 4A. The incidental training phase was also identical to Experiment 4A, except that participants received six blocks of training, rather than three. The test phase was identical to Experiment 4A, except that each block contained ten, rather than sixteen, stimuli. The number of test blocks was increased from eight to thirteen in order to approximately match the lengths of the test phases in Experiment 4A (128 trials) and Experiment 4B (130 trials).

Data archiving. The trial-level raw data is archived at `www.willslab.co.uk/plym6` with md5 checksum `bea925e24053a5e817c9950a8f2944fb`

3.6.2. Results and Discussion

In the intentional condition 32 of 35 participants passed the learning criterion (91%); for the incidental condition, the figure was 68% (27 of 40). Thus, our extension of the incidental training phase was successful in reducing the difference in proportion of learners across the two conditions, but it stopped short of equating the two conditions on this variable. We note that the pass rates for both conditions in the current experiment were similar to those of Kemler Nelson (1984, Experiment 1).

The top two rows of Table 8 show the results of both the traditional and the response-set analysis, after applying the learning criterion. Under the traditional analysis, the proportion of "overall similarity" responses to the critical test items was higher in the incidental condition than

		Traditional		Response-set		
	Condition	OS	CA	OS	CA	NCA
Passed	Intentional	.35	.65	.09	.63	.28
	Incidental	.57	.43	.15	.33	.52
ALL	Intentional	.37	.63	.09	.63	.28
	Incidental	.57	.43	.14	.30	.57

Table 8: Traditional and response-set analyses of Experiment 4B, for participants passing the learning criterion (Passed), and for all participants (ALL). Traditional: Proportion of Overall Similarity (OS) and Criterial Attribute (CA) responses to the critical test stimuli. Response-set: Proportion of participants best fit by an Overall Similarity (OS), Criterial Attribute (CA), and Non-criterial Attribute (NCA) response model.

in the intentional condition, $t(57) = 2.06, p = .04$. Thus Experiment 4B, like Experiment 4A, replicates Kemler Nelson's central result. Turning to the response-set analyses, Overall Similarity responders were rare, albeit slightly less so than in Experiment 4A. Thus, if the inclusion of overall-similarity-ambiguous test items in Experiment 4A had the effect of discouraging people from responding on the basis of overall similarity, this effect was presumably rather small. The current experiment reveals a trend for the proportion of Overall Similarity responders to be higher under incidental training than under intentional training, but this trend does not approach significance, $\chi^2(1) < 1$. The main result of Experiment 4B is that incidental training reduces the proportion of Criterial Attribute responders, $\chi^2(1) = 4.98, p = .03$. Non-criterial attribute responding is more prevalent under incidental training than under intentional training, and this difference is marginally significant, $\chi^2(1) = 3.47, p = .06$.

An analysis including all participants (Table 8, bottom two rows) leads to the same conclusions. Overall Similarity responding is rare, and the trend for a higher proportion of Overall Similarity responders under incidental conditions does not approach significance, $\chi^2(1) < 1$. Non-criterial Attribute responding is more prevalent under incidental conditions than intentional conditions, $\chi^2(1) = 5.72, p = .02$, and the reverse is true for Criterial Attribute responding, $\chi^2(1) = 7.45, p = .006$. The traditional analysis also remains significant, $t(73) = 2.26, p = .03$.

In summary, Experiment 4B supports the conclusions of Experiment 4A. In Experiment 5, we continued our re-examination of the criterial-attribute procedure, but turned to a different

independent variable — the effects of concurrent load.

3.7. Experiment 5

Smith and Shapiro (1989) employed the same abstract stimulus structure, basic procedures, and learning criterion, as the intentional condition of Kemler Nelson (1984), but with the caricatured face stimuli replaced by the visual presentation of pronounceable nonwords (e.g. BUNO), and with feedback learning replaced by observational learning (i.e. intentional training where the category label and stimulus are presented simultaneously). Some participants completed the training phase while also performing a secondary task (counting backwards in sevens), others completed it under full attention. All participants were then immediately transferred to a full-attention test phase. Participants trained under concurrent load were more likely to classify the critical test stimuli in an overall-similarity-consistent manner than participants trained under full attention. In the current experiment, we conducted a large-scale replication of the full-attention and concurrent load conditions of Smith and Shapiro (1989, Experiment 1), followed immediately by an extended test phase. Given that the presence or absence of OS-ambiguous test stimuli seemed to have little effect on the conclusions drawn from an incidental-intentional manipulation (compare Experiments 4A and 4B), and given the potential concern that excluding these stimuli might reduce sensitivity to distinguish criterial-attribute from non-criterial attribute strategies (see introduction to Experiment 4B), Experiment 5's test phase contained all 16 possible test stimuli.

3.7.1. Method

Participants, apparatus, and stimuli. 82 undergraduate psychology students at the University of Plymouth participated for course credit. The apparatus was identical to Experiment 4, with the addition of headphones for participants in the concurrent load condition. The abstract stimulus structure of Table 6 was instantiated using the same sets of pronounceable nonwords as Smith and Shapiro (1989, Experiment 1). Four different sets were used across participants: MUFA-VOSY, BUNO-KYPA, GIRU-LETA and DAKI-SEGO. For any given set, each letter position can take one of two values for a total of sixteen possible stimuli, all pronounceable. For example, the MUFA-VOSY set has the following sixteen stimuli: MUFA, MUFY, MUSA, MUSY, MOFA, MOFY, MOSA, MOSY, VUFA, VUFY, VUSA, VUSY, VOFA, VOFY, VOSA, VOSY

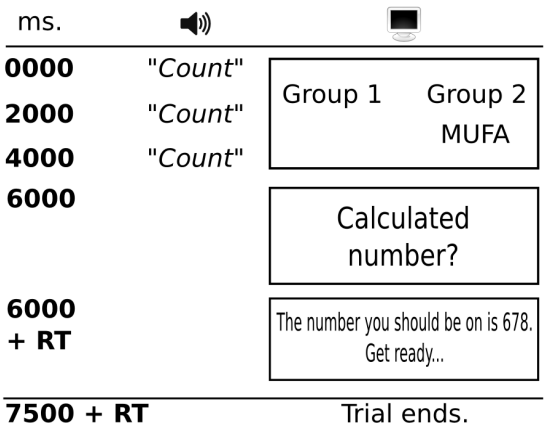

Figure 6: Example training trial in the concurrent load condition of Experiment 5. The left column shows time elapsed since the beginning of the trial, in milliseconds. RT = Participant's response time. The middle column represents what participants hear over headphones. The right column illustrates what is displayed on the monitor. Visual stimuli remain on the screen until replaced by the next visual stimulus.

Procedure. Participants were arbitrarily assigned to either concurrent-load training or full-attention training. In both cases, training was followed immediately by a full-attention test phase. Selection of the physical stimulus set (e.g. GIRU-LETA) was randomized across participants, as was the allocation of the criterial-attribute dimension (D1 in Table 6) to a specific letter pair within that set (e.g. R-T if the third letter pair of GIRU-LETA was selected), for a total of sixteen subconditions. Each subcondition was associated with a different allocation of the feature levels of each dimension to the categories (category 1 or category 2).

Concurrent-load training. Participants began with eight practice trials on the counting task alone; these trials were identical to the training phase (see below) except that the nonsense words and labels were not presented. Participants were then instructed that their task was to learn whether each of a series of nonsense words belonged to Group 1 or Group 2, while continuing to count backwards in sevens. They were also informed that Group 1 words would always appear to the left of the screen and Group 2 words would always appear to the right of the screen. Participants were assigned a random number between 800 and 900, and were instructed to repeatedly subtract 7 from that number each time they heard the word "count" (and only when they heard the word "count").

Figure 6 illustrates a typical trial in the concurrent-load condition. Each trial began with the

presentation of the label "Group 1" toward the left of the screen and the label "Group 2" toward the right of the screen. Simultaneously, a nonsense word was presented under the appropriate label, and the word "count" was presented over headphones. The word "count" was repeated at 2000 ms and 4000 ms after trial onset. At 6000 ms after stimulus onset, the nonsense word and category labels were replaced with the text "Calculated number?" Participants had 4000 ms from the onset of this text to enter their three-digit answer. After 4000 ms, or after they had entered three digits (whichever came first), the computer displayed the correct answer for the counting task (e.g. "The number you should be on is 855. Get ready...") for 1500 ms. If the participant had not entered any digits, the message "No response detected, please respond faster" was also displayed along with this correct answer. The training phase contained three blocks, each block comprising one presentation of each of the eight training stimuli listed in Table 6, in random order.

Full-attention training. The procedure was the same as in the concurrent-load condition, except that the counting task, and all mention of it, were removed. Hence, there were no counting-practice trials, the word "count" was not presented over headphones, the message "Calculated number?", and the requirement to enter a three-digit number, were removed, and the "number you should be on..." message was replaced with just "Get ready...". The "get ready" message was presented for 5500 ms in the full-attention training condition in order to approximately match the overall trial duration with the concurrent load condition (where participants had 4000 ms to enter their three-digit counting response). In practice, trial duration is likely to be somewhat shorter in the load condition, as participants need to respond before 4000 ms in order to avoid a time out message. This seems likely to further reduce the cognitive resources available in the concurrent load condition, relative to the full-attention condition, which is consistent with the main intention of the manipulation.

Test phase. Each trial commenced with the presentation of a nonword in the center of the screen for 2000 ms; any responses made during this period were ignored. After 2000 ms, the text "C - Group 1, M - Group 2" appeared under the nonword as a cue to respond and a reminder of the response keys. Participants then had a further 2000 ms to categorize the presented item, and were instructed to respond faster if a response was not detected. A 1500 ms inter-trial interval preceded the beginning of the next trial. No feedback or other information about category membership was given during the test phase. The test phase contained eight blocks, with each block comprising one

presentation, in random order, of all sixteen stimuli.

Data archiving. The trial-level raw data is archived at `www.willslab.co.uk/plym7` with md5 checksum `8962a66debd8178d6b74ade5e87fc71d`

3.7.2. Results and Discussion

In the full-attention condition, 90% (38 out of 42) of participants passed the learning criterion; for the concurrent-load condition, the figure was 55% (22 out of 40). Smith and Shapiro (1989) also reported lower pass rates under concurrent load than under full attention, although the difference in the current study is larger, and comparable to the difference in pass rates in Experiment 4A. One further participant was excluded from the response-set analysis on the grounds that their best-fitting model was a key bias (pressing the same key on every trial).

		Traditional			Response-set		
	Condition	OS	CA		OS	CA	NCA
Passed	Concurrent load	.53	.47		.00	.48	.52
	Full attention	.27	.73		.00	.74	.26
ALL	Concurrent load	.42	.58	*All-items*	.00	.41	.59
				10-items	.00	.46	.54
	Full attention	.31	.69	*All-items*	.00	.68	.32
				10-items	.02	.66	.32

Table 9: Traditional and response-set analyses of Experiment 5, for participants passing the learning criterion (Passed), and for all participants (ALL). Traditional: Proportion of Overall Similarity (OS) and Criterial Attribute (CA) responses to the critical test stimuli. Response-set: Proportion of participants best fit by an Overall Similarity (OS), Criterial Attribute (CA), and Non-criterial Attribute (NCA) response model, along with mean proportion of responses predicted by the best-fitting model (Consist.). All-participants response-set analysis is reported for all test items (All-items), and for the 10 test items for which the OS response model makes a clear prediction (10-items).

The top two rows of Table 9 show the results of both the traditional and response-set analysis of these data. Under the traditional analysis, the proportion of "overall similarity" responses to the critical test stimuli was higher in the concurrent-load condition than in the full-attention condition, $t(58) = 2.46, p = .02$. This replicates Smith and Shapiro's central result. Turning to the response-set analyses, no participant in either condition was best fit by an Overall Similarity response model.

The effect of concurrent load in the current analysis is thus wholly attributed to an increase in the proportion of Non-criterial Attribute responders, and a corresponding decrease in the proportion of Criterial Attribute responders, $\chi^2(1) = 4.01, p = .045$.

An analysis including all participants (see Table 9, bottom half, rows labelled "All-items") leads to similar conclusions. Under the traditional analysis, the proportion of "overall similarity" responses is higher in the concurrent-load condition than the full-attention condition, although this difference is not significant, $t(80) = 1.26, p = .21$. Given the now well-rehearsed problems of this traditional analysis, and the fact that Smith and Shapiro (1989) did not report an all-participants analysis, this failure of statistical significance is of relatively little consequence. More importantly, the response-set analysis once again reveals the absence of Overall Similarity responders in either condition, with the effect of concurrent load again fully attributed to an increase in the proportion of Non-criterial Attribute responders, with a corresponding reduction in the proportion of Criterial Attribute responders, $\chi^2(1) = 5.59, p = .02$.

Finally, a response-set analysis of all participants, restricted to the ten test stimuli for which an Overall Similarity strategy makes a clear prediction, leads to similar conclusions, albeit with only marginal significance. The proportions are shown in the bottom half of Table 9, in the rows labelled "10-items". A single participant in the full-attention condition is now best-fit by an Overall Similarity response model. The main effect of concurrent load is, again, an increase in the proportion of Non-criterial Attribute responders, $\chi^2(1) = 3.24, p = .07$, and a decrease in the proportion of Criterial Attribute responders, $\chi^2(1) = 2.54, p = .11$.

3.8. Interim summary

Taking the results of Experiments 4A, 4B, and 5 together, the conclusions are straight forward. Incidental training, and concurrent load, make it harder for participants to identify the one stimulus attribute that perfectly predicts category membership (the criterial attribute), from among three other attributes that predict membership with 75% accuracy (the non-criterial attributes). However, this increased difficulty does not significantly increase the likelihood that participants will employ an overall similarity strategy. Rather, it simply increases the likelihood that the single attribute they use is not the criterial attribute. This result can be accommodated within Combination Theory and thus, contrary to the conclusions of Kemler Nelson (1984) and Smith and Shapiro (1989), the effects of incidental training and concurrent load on the criterial-attribute procedure

do not favor Differentiation Theory over Combination Theory.

4. General Discussion

In the current paper we considered two, approximately opposite, theories about the order of processing in classification—Differentiation Theory and Combination Theory. Differentiation Theory assumes that classification starts with an undifferentiated whole, which can be separated into its constituent attributes if time and cognitive resources allow. In contrast, Combination Theory assumes classification starts with the attributes, and that information from these attributes can be combined and weighted if time and cognitive resources allow. A decade of research in our laboratory supporting Combination Theory over Differentiation Theory (Milton & Wills, 2004; Milton et al., 2008, 2009; Wills et al., 2013a,b) led us to re-examine a series of classic results using different procedures that are widely considered to support Differentiation Theory over Combination Theory (Kemler Nelson, 1984; Smith & Kemler Nelson, 1984; Smith & Shapiro, 1989; Ward, 1983). Taking the results of the seven experiments in the current paper together, the conclusions are straightforward — Combination Theory is favored over Differentiation Theory. This is because some results disconfirmed Differentiation Theory, and the remaining results were consistent with both theories. No experiment in this series was inconsistent with Combination Theory. Thus, overall, Combination Theory provided a better account of these data than Differentiation Theory. Indeed, Combination Theory was able to provide a complete account of the data across all seven experiments. In contrast, it seems unlikely that any version of Differentiation Theory could accommodate the results of Experiments 1 and 2, where the proportion of people consistently responding on the basis of one dimension and ignoring the other increases as stimulus presentation time reduces. According to Differentiation Theory, single-dimension responding requires analysis of the stimulus into its components, something that should be less, not more, likely as stimulus presentation time reduces.

4.1. Summary of findings

In Experiments 1, 2, 3A, and 3B, we re-examined the effects of time pressure on classification within the triad procedure (Smith & Kemler Nelson, 1984; Ward, 1983). The conclusion of previous studies had been that time pressure increases the prevalence of overall similarity classification; a result that has been taken to support Differentiation Theory over Combination Theory. However, traditional analysis of data from the triad procedure confounds classification on the basis of overall

similarity with classification on the basis of a single attribute (e.g. classifying on the basis of brightness, ignoring size). This confound can be overcome by employing the response-set analysis described in the current paper. In Experiment 1 (a re-analysis of Experiment 5 of Milton et al., 2008), we found that — in opposition to the conclusions of the traditional analysis — time pressure increased the proportion of participants responding on the basis of a single attribute. This result supports Combination Theory over Differentiation Theory. In Experiment 2, we ran a large-scale conceptual replication of Smith and Kemler Nelson (1984, Experiment 3), which also supported Combination Theory over Differentiation Theory, again opposite to the conclusions drawn from the confounded traditional analysis. In Experiments 3A and 3B, we ran two large-scale conceptual replications of Ward (1983, Experiment 2). These experiments produced a pattern of results consistent with both Combination Theory and Differentiation Theory. In summary, across our first four experiments, two were inconsistent with Differentiation Theory, but none were inconsistent with Combination Theory. We thus conclude, contrary to previous research, that the effects of time pressure on the triad procedure are better captured by Combination Theory than by Differentiation Theory.

In Experiments 4A, 4B, and 5, we re-examined the effects of incidental training and concurrent load on category learning within the criterial-attribute procedure (Kemler Nelson, 1984; Smith & Shapiro, 1989). The conclusion of previous studies had been that incidental training, and concurrent load, increase the prevalence of overall similarity classification; a result that has been taken to support Differentiation Theory over Combination Theory. However, traditional analyses of the criterial-attribute procedure have not fully distinguished overall similarity classification from classification on the basis of a single, non-criterial, attribute. In Experiments 4A and 4B, we applied response-set analysis to large-scale replications of the incidental vs. intentional training manipulation of Kemler Nelson (1984, Experiment 1). In both cases, the effect of incidental training, relative to intentional training, was to increase the proportion of participants classifying on the basis of a single, non-criterial attribute. There was no significant effect of incidental training on the proportion of Overall Similarity responders in either experiment. In Experiment 5, we applied the same response-set analysis to a large-scale replication of the concurrent load manipulation of Smith and Shapiro (1989, Experiment 1). The effect of concurrent load during training, relative to full-attention training, was again to increase the proportion of participants classifying on the

basis of a single non-criterial attribute. There was no significant effect of concurrent load on the proportion of Overall Similarity responders. We conclude, contrary to previous research, that the effects of incidental training and concurrent load on the criterial-attribute procedure are consistent with Combination Theory.

It is tempting to draw further inferences from the absolute proportions of Overall Similarity responders in the "low cognitive resource" conditions of Experiments 1-5. For example, one might argue that the rarity of Overall Similarity responders under incidental training (Experiment 4A) or concurrent load (Experiment 5) is surprising from the perspective of Differentiation Theory, because Overall Similarity responding is expected to be the least effortful strategy. One might also argue that the low prevalence of Unidimensional responding under time pressure in Experiment 3B is surprising from the perspective of Combination Theory, given Unidimensional responding is expected to be the least effortful strategy under this theory. However, on the basis of the current data, both inferences would be somewhat premature. Inspection of the low time-pressure conditions of Experiments 1-3 reveals wide variance in the baseline proportion of Overall Similarity responders in these studies. For example, there are no Overall Similarity responders in the 5000 ms condition of Experiment 2, but 56% of responders are best fit by an Overall Similarity model in the $>$ 5000 ms condition of Experiment 3B. A likely explanation for these differences is that some stimulus types evoke more overall similarity responding than others. The question of what aspects of the stimulus influence the prevalence of particular classification strategies is an interesting one (see Milton & Wills, 2004), but it is largely orthogonal to the current investigation. For current purposes, it is worth noting that Differentiation Theory is not favored over Combination Theory across a wide range of baseline rates of Overall Similarity classification.

Finally, although further research is needed, there are some indications that Overall Similarity classification is rather hard to elicit under any conditions in the criterial-attribute procedure. Across Experiments 4A, 4B, and 5, both caricatured faces and nonwords lead to very low rates of Overall Similarity responding, under both low- and high- cognitive resource conditions. If the rarity of Overall Similarity responding in the criterial-attribute procedure were to be replicated across a broader range of stimuli, this might be rather surprising from the perspective of Differentiation Theory, as Overall similarity classification is conceptualized as a low-effort strategy under this account.

4.2. Limitations and extensions

In the current series of experiments, we focussed on large-scale replications of four classic experiments: 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). A reasonable question is whether the outcome of our investigations would have been different if we had selected different experiments from within these multi-experiment papers. This question is considered in detail in the Supplementary Materials. In summary, the results of the other experiments in these papers pose no major problems for Combination Theory, and Combination Theory resolves two long-standing puzzles in this literature.

The first puzzle was why Kemler Nelson (1984, Experiment 2) appears to support Differentiation Theory while the very similar study of Ward and Scott (1987, Experiment 3) appears to support Combination Theory. There has been some debate between the authors of these two papers which, as far as we are aware, was never fully resolved (Kemler Nelson, 1988; Ward, 1988). Re-analysis on the basis of information available in the published literature demonstrates that Kemler Nelson (1984, Experiment 2) is, in fact, consistent with Combination Theory, thus resolving the controversy.

The second long-standing puzzle was why time pressure and concurrent load appear to have opposite effects in the experiments of Smith and Shapiro (1989). Combination Theory and response-set analysis provide the answer that both manipulations have the same effect — they both increase the prevalence of classifying on the basis of a single, arbitrarily chosen dimension. Thus, Smith and Shapiro's explanation of their time pressure result as people choosing "any attribute in a cognitive storm" (p. 394–395) may have wide applicability; it provides a colorful metaphor for what happens with both concurrent load and time pressure in both spontaneous classification and in category learning.

Beyond the four classic papers that were the focus of our investigations (Kemler Nelson, 1984; Smith & Kemler Nelson, 1984; Smith & Shapiro, 1989; Ward, 1983), a wider question is whether, if we had focussed on different papers or procedures, our conclusions would have been different. For reasons of length, this is not a question that can be addressed comprehensively in the current article, but a few illustrative cases are discussed below.

Integral stimuli. All stimuli in the current experiments seem likely to be separable rather than integral (Garner, 1976). The classification of stimuli as separable or integral is operational and includes

the property that separable stimuli are those for which selective attention to one of the attributes is relatively easy. There is some evidence that integral stimuli are nevertheless componential. For example, stimuli varying in pitch and loudness are integral, yet classification on the basis of pitch ignoring loudness is faster than a 45 degree rotation of that classification in stimulus space (Grau & Kemler Nelson, 1988). Similar results pertain for stimuli varying in saturation and brightness (Foard & Kemler Nelson, 1984). Thus, it is reasonable to ask what effects time pressure or concurrent load would have on the classification of integral stimuli. For example if, as seems likely, overall similarity classification truly dominates in an unspeeded triad task using hue-saturation stimuli, what would be the effect of reducing stimulus presentation time? Differentiation Theory would be disconfirmed by an increase in the prevalence of unidimensional classification in the face of reduced stimulus presentation time.

Incidental training. Love (2002) reported that, under incidental conditions, a particular type of overall similarity classification, a Type IV problem (Shepard et al., 1961), was easier to learn than an exclusive-or classification in which one stimulus dimension was irrelevant (a Type II problem). Under intentional conditions, Love found no significant difference between Type II and Type IV problems. One possible explanation of this result within the framework of Combination Theory is that incidental conditions promote simple additive combination rules (all that is required for Type IV) over more esoteric combination rules, such as exclusive-or. However, given a recent report that, even under intentional conditions, Type IV is sometimes easier than Type II (Kurtz et al., 2013), further research is probably required.

Feature-feature correlations. Combination Theory predicts that people should be more sensitive to feature-feature correlations under intentional conditions than incidental conditions. Differentiation Theory makes the opposite prediction. The results of Wattenmaker (1993, Experiment 1) support Combination Theory's prediction.

Developmental research. Both the triad procedure and the criterial-attribute procedure have been used to argue for a developmental shift away from overall similarity classification to more "dimensional" strategies (Kemler, 1982; Kemler Nelson, 1984; Minda & Miles, 2009; Smith & Kemler, 1977; Ward, 1980, 1983). However, this conclusion has been substantially critiqued (Ward & Scott, 1987; Ward et al., 1990; Thompson, 1994; Raijmakers et al., 2004), and the conclusions of these

critiques are uniformly compatible with Combination Theory.

Comparative research. Couchman et al. (2010) employed the criterial-attribute procedure in research comparing category learning in humans and monkeys. Their central conclusion was that humans, under intentional conditions, favor Criterial Attribute responding, while monkeys favor Overall Similarity responding. Their conclusion is quite different to the conclusion of Wills et al. (2009), who reported that, in a variant of the match-to-standards task, pigeons, squirrels, and humans all showed similar levels of Overall Similarity responding. In addition, Nicholls, Ryan, Bryant, and Lea (2011) reported that, in a slight variation of the criterial-attribute procedure, pigeons responded predominately on the basis of the criterial attribute. One possible reconciliation of these results is that pigeons, squirrels, monkeys and humans are, in fact, equally likely to use a single-attribute strategy, but humans are sometimes better at detecting and using the criterial attribute. Further research would be required to test the accuracy of this explanation.

4.3. Implications for a "family resemblance" account of natural categories

The idea that overall similarity classification requires substantial cognitive resources may seem paradoxical, given that natural categories appear to have an overall similarity structure (Mervis & Rosch, 1981; Rosch & Mervis, 1975), and are classified accurately and rapidly (Thorpe & Imbert, 1989). However, the exposure we receive to natural objects, and the practice we have in classifying them, presumably both contribute to the speed and accuracy with which everyday overall similarity categories are employed. It is an open question whether overall similarity classification would still be more effortful than single-dimension classification after thousands of trials of practice. By analogy with studies of automaticity of other cognitive skills (e.g. Logan, 1988), one possibility is that an initial effortful process of deriving overall similarity classifications would later be overtaken by fast retrievals from a store of previous exemplars.

4.4. Theory flexibility

Can the greater success of Combination Theory, relative to Differentiation Theory, be attributed to the former being a more flexible theory? Theory flexibility can become a problem when no result of an experiment could disconfirm the theory it was designed to test. Another way of expressing the same concern is to say that if a theory "predicts" everything that happens but also everything that doesn't happen, it is not a good theory. As we have made clear throughout the current paper,

there were things that could have happened in these experiments that would have disconfirmed Combination Theory. It's just that those things did not happen.

A more nuanced conceptualization of theory flexibility is to ask what proportion of potential outcomes a theory can accommodate. In this sense, Combination Theory is slightly more flexible than Differentiation Theory, but necessarily less flexible than a Combination-Differentiation hybrid theory. As Differentiation Theory cannot accommodate the full set of results in the current paper, Combination Theory appears to offer the best trade-off between adequacy and flexibility in providing an account of the results of the current paper.

4.5. Why combine?

If, as Combination Theory predicts, overall similarity classification is more effortful than single-dimension classification, then why is overall similarity classification ever observed in the triad and criterial-attribute tasks? Classifying on the basis of a single-dimension meets the stated task requirements, so why do people sometimes expend the extra effort? One possibility is that people prefer overall similarity classification strategies because they don't ignore noticeable variation in the stimulus dimensions. There could be a number of reasons for this preference. First, one might note that, outside the laboratory, most categories have an approximately overall similarity structure (Rosch & Mervis, 1975). Second, as Murphy (2002) pointed out, family resemblance categories are inductively rich, while single-dimension categories are inductively impoverished. In other words, if you observe a subset of the features of a family resemblance category, you can (imperfectly) induce unobserved features. For single-dimension categories, you need to observe the single feature to classify the item, and the category has no other characteristic features, so there are no further inductions that can be drawn.

4.6. Combination Theory: Concluding remarks

Combination Theory assumes that classification starts with the stimulus attributes, and that information from these attributes can be combined and weighted if time and cognitive resources allow. Over the last decade, a body of work from our laboratory has substantially increased the evidence base for Combination Theory (Milton & Wills, 2004; Milton et al., 2008, 2009; Wills et al., 2013b,a). However, nearly all our previous studies in this area have used a single procedure (the match-to-standards procedure). Evidence from two other procedures — the triad procedure (Smith

& Kemler Nelson, 1984; Ward, 1983) and the criterial-attribute procedure (Kemler Nelson, 1984; Smith & Shapiro, 1989) — appeared to support Differentiation Theory, an approximately opposite theory. One understandable response to this lack of agreement would be to argue that the match-to-standards procedure is in some way methodologically flawed or atypical. As a result of the current investigations, it seems more likely that the apparently opposite conclusions were an artifact of the confounded analysis method employed in triad task and criterial-attribute investigations. Match-to-standard investigations have, from the outset, used the less confounded response-set analysis.

In conclusion, when better methods of analysis are employed, the match-to-standards, triad, and criterial-attribute, procedures form a coherent body of work that either supports Combination Theory over Differentiation Theory, or is consistent with both theories.

Acknowledgments

This research was supported by ESRC grant RES-000-22-1779 awarded to Andy J. Wills. The authors thank Jacob Brain, Anna Kharko, Chris Longmore, Martin O'Shaughnessy, and Daniel Perkins for their assistance with data collection.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch, & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Erlbaum.

Cohen, A. L., & Nosofsky, R. M. (2003). An extension of the exemplar-based random-walk model to separable-dimension stimuli. *Journal of Mathematical Psychology*, *47*, 150–165.

Couchman, J. J., Coutinho, M. V. C., & Smith, J. D. (2010). Rules and resemblance: Their changing balance in the category learning of humans (*Homo sapiens*) and monkeys (*Macaca mulatta*). *Journal of Experimental Psychology: Animal Behavior Processes*, *36*, 172–183.

Foard, C. F., & Kemler Nelson, D. G. (1984). Holistic and analytic modes of processing: The multiple determinants of perceptual analysis. *Journal of Experimental Psychology: General*, *113*, 94–111.

Garner, W. R. (1976). Interaction of stimulus dimensions in concept and choice processes. *Cognitive Psychology*, *123*, 98–123.

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, *65*, 231–262.

Grau, J. W., & Kemler Nelson, D. G. (1988). The distinction between integral and separable dimensions: Evidence for the integrality of pitch and loudness. *Journal of Experimental Psychology: General*, *117*, 347–70.

Gross, C. G. (2008). Single neuron studies of inferior temporal cortex. *Neuropsychologia*, *46*, 841–852.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, *148*, 574–591.

Jeffreys, H. (1961). *The Theory of Probability*. (3rd ed.). Oxford: Oxford University Press.

Kemler, D. G. (1982). Classification in young and retarded children: the primacy of overall similarity relations. *Child Development*, *53*, 768–79.

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.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 552–72. doi:10.1037/a0029178.

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General*, *124*, 161–180.

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 695–711.

Larsen, A., & Bundesen, C. (1992). The efficiency of holistic template matching in the recognition of unconstrained handwritten digits. *Psychological Research*, *54*, 187–193.

Lockhead, G. R. (1972). Processing dimensional stimuli: A note. *Psychological Review*, *79*, 410–419.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*, 829–35.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*, 242–279.

Mervis, C. B., & Rosch, E. (1981). Catgeorization of natural objects. *Annual Review of Psychology*, *32*, 89–115.

Milton, F., Longmore, C. A., & Wills, A. J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 676–692.

Milton, F., Viika, L., Henderson, H., & Wills, A. (2011). The effect of time pressure and the spatial integration of the stimulus dimensions on overall similarity categorization. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 795–800). Austin, TX: Cognitive Science Society.

Milton, F., & Wills, A. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 407–415.

Milton, F., & Wills, A. (2009). Eye movements in overall similarity and single-dimension sorting. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1512–1517). Austin, TX: Cognitive Science Society.

Milton, F., Wills, A. J., & Hodgson, T. L. (2009). The neural basis of overall similarity and single-dimension sorting. *NeuroImage*, *46*, 319–326.

Minda, J. P., Desroches, A. S., & Church, B. A. (2008). Learning rule-described and non-rule-described categories: A comparison of children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1518–1533.

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.

Murphy, G. L. (2002). *The Big Book of Concpets*. Cambridge, MA: MIT Press.

Neisser, U. (1967). *Cognitive Psychology*. Englewood Cliffs, NJ: Prentice Hall.

Newell, B. R., Dunn, J. C., & Kalish, M. (2010). The dimensionality of perceptual category learning: a state-trace analysis. *Memory & Cognition*, *38*, 563–581.

Nicholls, E., Ryan, C. M. E., Bryant, C. M. L., & Lea, S. E. G. (2011). Labeling and family resemblance in the discrimination of polymorphous categories by pigeons. *Animal Cognition*, *14*, 21–34.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104–114.

Nosofsky, R. M., & Kruschke, J. K. (2002). Single-system models and interference in category learning: Commentary on Waldron and Ashby (2001). *Psychonomic Bulletin & Review*, *9*, 169–174.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*, 172–191.

Overstall, A. M. (2014). *conting: Bayesian analysis of contingency tables (R package version 1.2)*. URL: `http://CRAN.R-project.org/package=conting`.

Overstall, A. M., & King, R. (2014). conting : An R package for Bayesian analysis of complete and incomplete contingency tables. *Journal of Statistical Software*, *58*, e1–27.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*, 341–365.

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.

Regehr, G., & Brooks, L. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: General*, *122*, 92–114.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*, Whole No. 517.

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.

Smith, L. B. (1989). A model of perceptual classification in children and adults. *Psychological Review*, *96*, 125–144.

Smith, L. B., & Kemler, D. G. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology*, *24*, 279–298.

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.

Thompson, L. A., & Massaro, D. W. (1989). Before you see it, you see its parts: Evidence for feature encoding and integration in preschool children and adults. *Cognitive Psychology*, *21*, 334–362.

Thorpe, S. J., & Imbert, M. (1989). Biological constraints on connectionist modeling. In Z. Pfeifer, F. Schreter, F. Fogelman-Soulie, & L. Steels (Eds.), *Connectionism in Perspective* (pp. 63–92). Amsterdam: Elsevier.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *136*, 97–136.

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: evidence for multiple category learning systems. *Psychonomic Bulletin & Review*, *8*, 168–176.

Ward, T. B. (1980). Separable and integral responding by children and adults to the dimensions of length and density. *Child Development*, *51*, 676–84.

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.

Ward, T. B., Vela, E., & Hass, S. D. (1990). Children and adults learn family-resemblance categories analytically. *Child Development*, *61*, 593–605.

Wattenmaker, W. D. (1993). Incidental concept learning, feature frequency, and correlated properties. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 203–22.

Wills, A. J., Lea, S. E. G., Leaver, L. A., Osthaus, B., Ryan, C. M. E., Suret, M. B., Bryant, C. M. L., Chapman, S. J. A., & Millar, L. (2009). A comparative analysis of the categorization of multidimensional stimuli: I. Unidimensional classification does not necessarily imply analytic processing: Evidence from pigeons (*Columba livia*), squirrels (*Scurius carolinensis*), and humans (*Homo sapiens*). *Journal of Comparative Psychology*, *123*, 391–405.

Wills, A. J., Longmore, C. A., & Milton, F. (2013a). Impulsivity and overall similarity classification. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 3783–3788). Austin, TX: Cognitive Science Society.

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

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34*, 387–398.