Absence of cross-modality analogical transfer in perceptual categorization

C. E. R. Edmunds*   Angus B. Inkster[†]   Peter M. Jones[†]   Fraser Milton[‡]   Andy J. Wills[†]

Abstract

Analogical transfer has been previously reported to occur between rule-based, but not information-integration, perceptual category structures (Casale, Roeder, & Ashby, 2012). The current study investigated whether a similar pattern of results would be observed in cross-modality transfer. Participants were trained on either a rule-based structure, or an information-integration structure, using visual stimuli. They were then tested on auditory stimuli that had the same underlying abstract category structure. Transfer performance was assessed relative to a control group who did not receive training on the visual stimuli. No cross-modality transfer was found, irrespective of the category structure employed.

Keywords: analogical reasoning, generalisation, strategies, learning, multiple systems

1   Introduction

Analogical transfer is the ability to transfer knowledge of structural features of a problem from one domain to another (Gentner & Maravilla, 2018). The canonical example (Gick & Holyoak, 1980) asked: "how do you destroy a tumor using intense radiation without damaging the surrounding tissue?" (Duncker, 1945). People found this easier to solve if they had previously experienced an analogous problem in a military context ("how can you storm a castle with a large army when the surrounding bridges are weak?"). Obviously, the worth of transferring knowledge extends far beyond satisfying the whims of experimental psychologists. For instance, it is central to scientific work, from hypothesis generation to data interpretation (Dunbar, 1997). Further, without analogical transfer much of the effectiveness of education is lost (Ford, Baldwin, & Prasad, 2018). After all, who would trust a nurse (Johnston, Coyer, & Nash, 2017) or engineer (Chase, Malkiewich, & Kumar, 2019) who is unable to generalise their training?

In the transfer literature, one rarely-explored task is perceptual categorization, where people learn to label groups of perceptually similar stimuli (Kurtz, 2015). For instance, a doctor might categorize scans as showing evidence of cancer or remission (Hornsby & Love, 2014). Analogical transfer research on perceptual categorization has solely examined unidimensional and information-integration tasks (see Figure 1; Ashby & Gott, 1988).[1] For ease of exposition, assume we are categorizing Gabor patches varying in the number and orientation of bars. In unidimensional (UD) tasks (see Figure 1A), participants sort stimuli using a single dimension: "Patches with bars orientated less than 45 degrees are in Category A; otherwise they are in Category B." In contrast, in information-integration (II) tasks (see Figure 1B) participants combine features from two incommensurable dimensions.[2] These two tasks are rotations of each other in stimulus space and so are argued to be well-matched on between-category distance, within-category similarity and difficulty (e.g. Smith, Boomer, et al., 2014; Ashby and Valentin, 2018; although see Newell, Dunn, and Kalish, 2011). Indeed, a large literature has been built on the assumption that the principal difference between these two structures is that they differ in how easy it is to describe the decision-boundary between the categories (Ashby & Maddox, 2005, 2011; Ashby & Valentin, 2017).

Despite their many similarities, these categorization tasks do not appear to be equally amenable to analogical transfer. In this context, analogical transfer would be shown "if a participant is able to apply a classification strategy learned with one set of stimuli to a set of novel, perceptually distinct stimuli" (Casale et al., 2012, p. 434). In Casale et al.'s (2012) experiments, participants learned to categorize Gabor patch stimuli according to a UD or II structure. To test the extent

*School of Biological and Chemical Sciences, Queen Mary, University of London, U.K. Email: ceredmunds@gmail.com

[†]School of Psychology, University of Plymouth, U.K.

[‡]School of Psychology, University of Exeter, U.K.

[1]There has also been work examining the transfer of relational categories which is outside the definition of perceptual categorization used here.

[2]II structures are difficult to describe. However, for those interested, a possible description of an II category in Gabor patch stimuli is "If there are more bars than the bars are upright then the stimulus is in Category A, otherwise it is in B." This description sounds nonsensical and so is unlikely to be identified by participants.
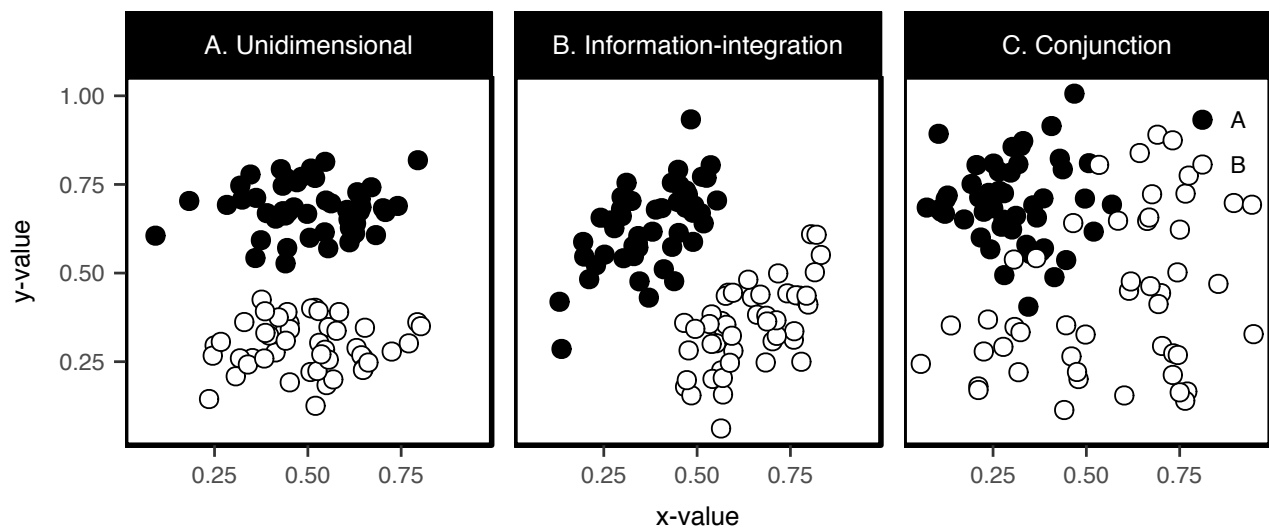
FIGURE 1: Examples of A. Unidimensional, B. Information-integration and C. conjunction category structures. Each circle represents a stimulus. Filled circles indicate stimuli belonging to Category A, and empty circles those belonging to Category B.

of analogical transfer, participants then completed similar categorization tasks with stimuli perceptually distinct from the training set. These novel test stimuli were chosen such that the between-category decision boundary was the same for both sets of stimuli. Across three experiments Casale et al. found that participants who learned a UD structure could transfer that knowledge to another stimulus space region. However, those who learned an II structure could not. In a similar study that used rectangles of visual static (random green dots) that varied in dot density and rectangle size, Zakrzewski, Church, and Smith (2018) found the same pattern (see also Soto & Ashby, 2019).

In summary, previous studies indicate that analogical transfer is possible for UD tasks but not for II tasks. However, these studies limit the inferences we can make about analogical transfer as they are all examples of so-called "near transfer," where both tasks are presented in a similar context, using similar stimuli (Barnett & Ceci, 2002). Of more interest is "far transfer," such as the radiation problem mentioned above, where the surface task features differ between training and test. After all, the more situations to which we can transfer knowledge, the more use it is. Therefore, in the current work we investigate far transfer in perceptual categorization.

1.1 The current work

We examined analogical transfer by giving participants a training phase followed by a test phase with no feedback (as in Experiment 3 of Casale et al., 2012). However, to investigate far transfer, we changed the surface features of the stimuli between training and test. The training stimuli were boxes intersected by a line that varied in its length

and height within the box (see Figure 2). The test stimuli were auditory tones that varied in length and pitch. We chose auditory tones as previous work showed successful learning of rule-based and II category structures using these stimuli (Maddox, Ing, & Lauritzen, 2006; Reetzke, Maddox, & Chandrasekaran, 2016; Smith, Johnston, et al., 2014; Yi, Maddox, Mumford, & Chandrasekaran, 2016). Successful analogical transfer required participants to understand that the length and height of the line is analogous to the length and pitch of the tone. Further, following previous work in both categorization (e.g., Edmunds, Milton, & Wills, 2015; Edmunds, Wills, & Milton, 2016, 2019; Edwards, Williams, Gentner, & Lombrozo, 2019) and analogical transfer (e.g., Gick & Holyoak, 1980), participants were asked to give verbal reports. We used these reports to examine whether participants had transferred explicit knowledge between the phases of the experiment.

In addition to changing the surface features of the stimuli, we exchanged the UD rule-based task for a conjunction (CJ) task (see Figure 1C). As noted above, UD and II tasks are well-matched on several features, such as the maximum accuracy achievable, with- and between-category similarity, category size and so on (e.g., Ashby & Valentin, 2018; Smith, Boomer, et al., 2014). However, they are not matched on the number of stimulus dimensions required to make categorization responses: UD tasks require attention to one dimension, II tasks to two. This difference in task complexity previously undermined some of the experiments using II category tasks (e.g., Carpenter, Wills, Benattayallah, & Milton, 2016; Edmunds et al., 2015). Therefore, we used a CJ task as a more appropriate control for the II task.

Finally, we also added a control task. In pilot work[3], we found that the strategies participants reported using differed from what we predicted. In previous studies, the majority of participants used the optimum diagonal strategy in the II task throughout the experiment (Casale et al., 2012; Soto & Ashby, 2019; Zakrzewski et al., 2018). In contrast, we found that the majority of participants reported using two-dimensional rules at training and one-dimensional rules at test.

One possible explanation is that participants misinterpreted the change of modality as signaling the start of an entirely new categorization task. Considered under that lens, the test phase would appear to participants as a simple unsupervised learning task. There is considerable evidence that shows that without feedback participants often resort to using unidimensional rules (e.g. Ashby, Queller, & Berretty, 1999; Milton, Longmore, & Wills, 2008; Wills, Inkster, & Milton, 2015). Further, using a unidimensional rule in II tasks results in performance around 75%, well above chance (Edmunds, Milton, & Wills, 2018). To test this possibility we added control conditions (one for each category task), where participants had to complete the final transfer test without prior categorization training. This provided us with a baseline to test improvement when given additional training.

2 Method

2.1 Participants

The participants were 117 undergraduate psychology students from the University of Plymouth who received partial course credit for their participation.

The sample size was calculated a priori using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) based on the effect size of previous work. From the most similar prior experiment (Experiment 3 Casale et al., 2012), we calculated η_p^2 as 0.27 for the key interaction term between performance across training and test and categorization task. This implied we needed at least 48 participants to achieve 80% power. Given that published work often over-estimates effect sizes (Button et al., 2013) and our task is harder so we might expect a smaller effect, we increased the number to around thirty per cell.

2.2 Category structures and stimuli

This experiment had a 2 (Category structure: Conjunction, Information-integration) x 2 (Task: Control, Analogical transfer) between-subjects design. The participants were randomly assigned to one of the four conditions.

There were two sets of stimuli: visual and auditory. The visual stimuli consisted of 96 monochromatic 'boxes' pre-
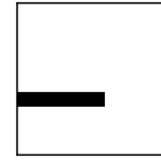

FIGURE 2: Example visual stimulus

sented on a white background (see Figure 2). The stimuli varied in the length and height of the interior line. This was calculated in pixels from the abstract coordinates (see below) using $f(x) = x/6 - 0.15$ for the length dimension and $f(x) = x/105$ for the line-height dimension. The width of the interior line (24 pixels) and the size of the outer square box (each side was 200 pixels long and 3 pixels thick) remained constant throughout the experiment.

The abstract coordinates are shown in Table 1. The auditory stimuli were 96 tones based on those used by Maddox et al. (2006), that varied in duration and pitch. They were narrow-band noise, created by combining 50 independent phase sine waves. The frequencies of the stimuli were drawn from a normal distribution with a mean value of F Hz and a standard deviation of 15.7. F were given by $F = 1000 * H + 150$ where H are the coordinates for height from the visual stimuli. The durations of the stimuli were given by the equation $D = 2 * (0.1 + 10^{L-1})$, where L are the arbitrary stimulus coordinates for length from the visual stimuli.

The category structures were also counterbalanced. For the CJ structure, this resulted in four structures equivalent to rotating the category structure in Table 1 by $\pi/2$ each time. There were 19 participants who assigned the top-left to Category A, 19 who assigned the top-right, 10 who assigned the bottom-left and 11 who assigned the bottom-right. For the II structure, there were two structures each a π rotation of the other. There were 28 participants who learned a category

TABLE 1: Abstract category structures

	Height		Length		
	μ	σ	μ	σ	cov_{xy}
Conjunction					
A	0.3	0.12	0.8	0.12	
B	0.3	0.12	0.4	0.12	
B	0.7	0.12	0.4	0.12	
B	0.7	0.12	0.8	0.12	
Information-integration					
A	3.1562	0.667	66.5	12.247	5
B	4.7546	0.667	37	12.247	5

---

[3]available at https://osf.io/mcd7v

structure with a positive correlation between dimensions and 30 that learned the rotation.

2.3　The questionnaire

After they had completed the computer portion of the task, each participant answered the questions reproduced in Table 2. Participants in the control condition completed the three relevant questions from the longer questionnaire.

We also asked two multiple choice questions about the visual and auditory tasks respectively. However, due to an error these were not interpretable and so have not been further considered.

Note that in previous experiments using these types of category structures, researchers have used a model-based decision-bound analysis informed by signal detection theory (Ashby & Gott, 1988; Maddox & Ashby, 1993). However, recent work has shown that this analysis can be unreliable and consequently hard to interpret (Donkin, Newell, Kalish, Dunn, & Nosofsky, 2015; Edmunds et al., 2018). Therefore, we have not included a description of this analysis in the main text, although it is available in the Appendix for those interested.

2.4　Materials

The experiment was run on a PC connected to a 22-inch monitor with a 16:9 aspect ratio using E-Prime 2.0. Responses were made using a standard keyboard. The auditory stimuli were presented on Behringer HPM1000 on-ear headphones.

2.5　Procedure

Participants in the analogical transfer conditions completed both the training and test phases, whereas participants in the control condition only completed the test phase. The specific stimuli each participant was shown depended on the condition they were assigned. The training phase consisted of four blocks of 96 training trials. Participants were told they would be shown some 'boxes' containing a single line and that their task was to identify the box as either Category A or Category B. Participants were also informed that the box stimuli would differ on two dimensions, height and length. Participants were told to indicate the category by pressing 'Z' for Category A and 'M' for Category B. Participants were told that they would have to guess which category each 'box' belonged to at the start but that through corrective feedback they could achieve high accuracy.

On each trial, the participants were presented with a box stimulus for 1000 ms, which approximately matched the average duration of the auditory stimuli. This was replaced with the question "Category A or Category B?" This remained onscreen for 1000 ms. Once a response was detected, feedback was displayed for 1000 ms: either "Correct" in blue,

or "Incorrect" in red as appropriate. If the participant failed to respond within 1000 ms, "TIME OUT" was displayed in red instead of feedback. The inter-trial interval was 500 ms. There were no signals to indicate the end or beginning of an experimental block.

The test phase consisted of a single block of 96 auditory stimuli. Participants were told that they would hear a series of sounds varying in duration and pitch. They were told that they would have to categorize the sounds by pressing "Z" for Category A or "M" for Category B. Participants did not receive feedback for this part of the experiment. As in Maddox et al. (2006), the loudness of the stimuli were not controlled using a normalisation procedure.

On each trial, participants were presented with an auditory stimulus for the length of time specific to the stimulus. Then, the question "Category A or Category B?" was presented for 1000 ms, during which time the participant was allowed to respond. After receiving a response, a blank screen was presented for 1500 ms, before the start of the next trial. However, if the participant failed to respond within 1000 ms, "TIME OUT" was displayed for 1000 ms followed by a blank screen for 500 ms. Before being debriefed, participants completed the short questionnaire about the experiment.

2.6　Analysis

All data analyses were conducted in R (R Core Team, 2017). Cohen's d was calculated using the `effsize` R package (Torchiano, 2017) and Bayes Factors using the `BayesFactor` package (Morey & Rouder, 2018). Bayes Factors (BF_{10}) exceeding three were considered evidence for the experimental hypothesis, while $BF_{10} < \frac{1}{3}$ were considered evidence for the null; p-values are largely reported as a traditional courtesy, except where no Bayesian equivalent was known to us, in which case an alpha level of .05 was adopted. We also report generalised η^2. Trial by trial data, verbal reports and all analyses for this and three prior pilot experiments are available at `https://osf.io/mcd7v`.

2.7　Results

2.7.1　Data rescoring

A Shapiro-Wilks test showed that accuracy in the test phase in all four conditions violated assumptions of normality, ($p \leq .001$). On visually inspecting the distributions of accuracy scores during the test phase, both conditions had distinctly bimodal distributions with peaks around 30% and 70%. Matching peaks 20% away from chance indicates that all participants were exhibiting systematic behavior related to the category structure, but were at chance on inferring which key went with which category label.

There are two ways of solving this problem. One way is to use non-parametric tests. The other is to rescore the data so

TABLE 2: Post-experiment verbal report questionnaire.

| Question	Type
1 Did you notice any relationship between the first part of the experiment (pictures) and the second part of the experiment (sounds)?	Yes/No
2 If your answer to Question 1 was YES, please describe the relationship below:	Open
3 For the first part of the experiment (pictures), are you able to describe how you decided whether a picture belonged in category A or category B?	Yes/No
4 If your answer to Question 3 was YES, please describe how you did it below:	Open
5 For the second part of the experiment (sounds), are you able to describe how you decided whether a sound belonged in category A or category B?	Yes/No
6 If your answer to Question 5 was YES, please describe how you did it below:	Open
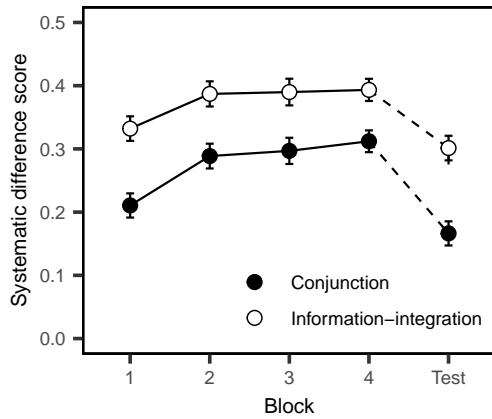
FIGURE 3: Learning curves of participants. Error bars are difference-adjusted between-subject 95% confidence intervals (Baguley, 2012).

that it reflects difference from chance. This is where we think category learning differs from other research on analogical transfer. In 'traditional' analogical transfer problems, such as the radiation problem, systematicity and accuracy are positively correlated. The more accurate you are, the better you have transferred the information from one task to another. In contrast, for a category learning task, systematicity forms a U-shaped curve on accuracy. Participants that score 100% are also perfectly systematic, but so are those that score 0%. Therefore, as we are interested in the 'systematicity' of participants responses we chose the later approach. The rescored performance metrics 'Systematic difference scores' are shown in Figure 3 for the transfer conditions and Figure 4 for the test block.

2.7.2 Transfer

We conducted a 2 (task: CJ, II) x 2 (block: last training block, test) ANOVA on the data from the full condition. There was a main effect of block, $BF_{10} = 1.4 \times 10^8, F(1, 57) = 142.04, \eta^2 = 0.42, p < .001$: participants scored lower in the test block than at the end of training. There was also a main effect of category task, $BF_{10} = 1.7 \times 10^6, F(1, 57) = 49.39, \eta^2 = 0.38, p < .001$: participants scored higher on average in the II condition, $M_{II} = 0.35, SD = 0.08$, than in the CJ condition, $M_{CJ} = 0.24, SD = 0.10$. The evidence for an interaction was equivocal, $BF_{10} = 1.61, F(1, 57) = 7.25, \eta^2 = 0.04, p = .009$. The numerical trend was for performance to drop more in the rule-based (CJ) condition than in the II condition; this is opposite to the pattern reported in previous work.

2.7.3 Comparison with controls

Here, we examined whether training had any impact on final test performance. A 2 (Category structure: CJ, II) x 2 (Condition: Control, Experimental) ANOVA on test block accuracy found equivocal evidence for an interaction, $BF_{10} = 0.43, F(1, 113) = 1.03, \eta^2 = 0.01, p = .311$, and evidence for the *absence* of a main effect of condition, $BF_{10} = 0.25, F(1, 113) = 1.07, \eta^2 = 0.01, p = .302$. There was a main effect of category structure, $BF_{10} = 5.2 \times 10^{11}, F(1, 113) = 79.22, \eta^2 = 0.41, p < .001$; participants scored higher on average on the II structure, $M_{II} = 0.29, SD = 0.07$, than on the CJ structure, $M_{CJ} = 0.17, SD = 0.08$.

2.8 Verbal reports

The open questions were coded by the first author (CERE) and a volunteer (MB). For the answer to Question 2, participants were identified by how they paired stimulus dimensions across the two phases of the experiment. For instance, "short line matches short tone" would have been coded as length-duration. The correct relationship was "line length is tone duration and line position is tone pitch."

For the answers to Questions 4 and 6, participants were identified as using one of the following strategies:
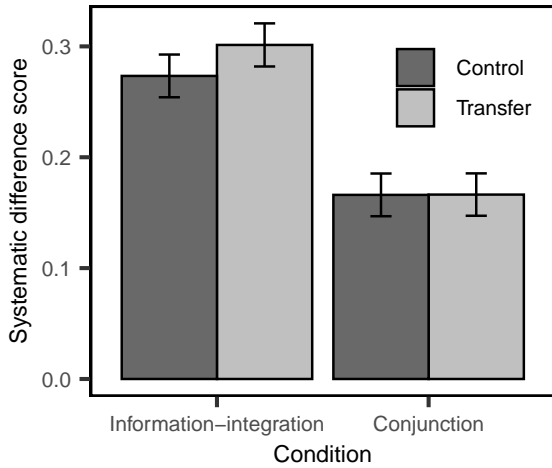
FIGURE 4: Performance on the auditory test stimuli. Error bars are difference-adjusted between-subject 95% confidence intervals (Baguley, 2012).

TABLE 3: Rows 1-3: Proportion of participants in the analogical transfer conditions responding 'yes' to verbal report questions 1, 3, and 5. Row 4: proportion of control participants responding 'yes' to their equivalent of question 5. The chi-squared test (and corresponding Bayes Factor) compare these proportions between categorization tasks.

	CJ	II	χ^2	BF_{10}
Q1. Relationship?	0.65	0.64	0.00	0.34
Q3. Strategy? (Training)	1.00	0.96	0.73	0.14
Q5. Strategy? (Transfer)	0.75	0.79	0.08	0.31
Q5. Strategy? (Controls)	0.86	1.00	4.26	1.09

CJ = Conjunction, II = information-integration.

TABLE 4: Proportion of participants reporting each strategy.

	Phase	UD	CJ	Complex	Other
				Strategies	
CJ	Training	0.05	0.80	0.15	-
	Test	0.47	0.40	0.13	-
	Control	0.83	-	0.07	0.06
II	Training	0.26	0.15	0.59	-
	Test	0.73	0.09	0.18	-
	Control	0.93	-	0.07	-

Strategies: UD=Unidimensional, CJ=Conjunction, Complex=Complex multi-dimensional

Participants were classified as using a *unidimensional* rule if they only mentioned a single stimulus dimension, such as "if the line was short the stimulus was in Category A, if the line was long the stimulus was in Category B."

Participants were classified as using a *conjunction* rule if they mentioned both stimulus dimensions and described categorizing stimuli on the basis of a logical "AND" rule such as "short stimuli in the top half of the box were in Category A, otherwise they were in Category B."

Participants were classified as using a *complex* rule if they mentioned both stimulus dimensions and described categorization stimuli on the basis of a series of complicated if statements, such as "If the line is short it is in Category A, but if it is long it is in Category B. However, if it is mid-length then if it is high in the box it is Category B, and if it is low in the box it is in Category A."

Participants were classified as *other* if they did not fall into one of the groups above. Inter-rater reliability was 82%. All discrepancies between raters were easily resolved following discussion.

Table 3 shows the proportion of participants that responded "yes" to the yes/no questions of Table 2, along with the corresponding statistical tests. There was evidence for the *absence* of a difference between the CJ and II conditions in the proportions of participants that reported a strategy in the training phase. The same was true for the test phase. The evidence was equivocal for a difference between CJ and II in the proportion of participants who reported noticing a relationship between the two phases of the experiment. In the control conditions, numerically more people were able to report a strategy in the II task than in the CJ task, but the Bayesian evidence for that difference was equivocal

Next, we looked at the strategies participants reported (see

Table 4). No participants attempted to report anything that could even be loosely interpreted as implicit responding. Rather, for participants completing the full experiment, the majority of participants reported shifting from using two-dimensional rules during training to unidimensional rules at test. In the CJ conditions, participants relied more on a unidimensional strategy in the control group than at test, $BF_{10} = 4.12$, $\chi^2 = 4.95$, $p = .026$. In contrast, there was no evidence that this was the case in the II conditions, $BF_{10} = 1.46$, $\chi^2 = 3.72$, $p = .054$. This interaction reached significance, $BF_{10} = 7.15$, $\chi^2 = 12.65$, $p = .013$. There was no evidence to suggest that participants relied less on the default unidimensional categorization strategy following CJ training than II training, $BF_{10} = 1.31$, $\chi^2 = 2.58$, $p = .109$.

Finally, we looked at the proportion of participants in the full experiment conditions that were able to correctly specify the correct transfer mapping. In the rule-based condition, 38% of participants correctly specified the transfer relationship, and 5% of those in the II condition. The Bayesian evidence for this difference was equivocal, $BF_{10} = 1.75$,

$\chi^2 = 3.81$, $p = .051$. No participants were identified as using a different mapping (such as height of bar to length of tone). In the CJ condition, participants who reported the correct mapping performed worse, $M = 0.14$, than those who reported an incorrect mapping, $M = 0.18$, although the evidence for this difference equivocal, $BF_{10} = 0.59$, $t(11) = −0.70$, $p = .501$.

3   Discussion

Analogical reasoning is crucial to transferring knowledge in many tasks and domains (Barnett & Ceci, 2002; Gentner & Maravilla, 2018). However, the role analogical reasoning plays in transferring category knowledge has only been partially explored. Some studies have explored analogical reasoning in near transfer, where the pre-transfer and post-transfer tasks are in the same context (Casale et al., 2012; Soto & Ashby, 2019; Zakrzewski et al., 2018). They found that successful analogical transfer was contingent on the category structure: transfer was easier following UD training than II training.

We extended the literature by investigating far transfer of category knowledge; in this case where category knowledge might have been transferred between two different modalities. Participants were trained on a visual task and tested on an auditory task with the same underlying structure. There was evidence that, compared to the relevant control conditions, fewer participants used a unidimensional strategy in the CJ conditions than the II conditions. However, unlike the studies of near transfer, this did not result in improved accuracy: participants who received training on the visual task achieved the same levels of accuracy in the auditory task as those who did not receive training, irrespective of category structure.

The current findings contrast those in reasoning tasks, where analogical reasoning improves performance in the second task despite substantial changes in surface form (Barnett & Ceci, 2002; Gentner & Maravilla, 2018). This contrast may be due to task differences. In both kinds of task, participants might identify the optimum strategy, yet have problems implementing it. However, these problems of implementation seem likely to be more severe for a categorization task than a classic reasoning task. For example, knowing that "small" stimuli are in one category and "large" stimuli are in another can be hard to apply effectively if you do not know how small is "small" and how large "large". It it possible that these fine-grained adjustments require feedback.

3.1   Learning versus default strategy use

In the control condition of the current experiment, classification of the auditory stimuli was performed in the absence of prior in-experiment training, and in the absence of feedback.

We found, as others have previously (Ashby et al., 1999; Milton et al., 2008; Wills et al., 2015), that people do not guess in such situations, they instead apply a default, typically single-dimensional, strategy. It is reasonable to assume that they might at least start with a similar default strategy even when feedback is available. The issue this creates is that applying a single-dimension strategy to CJ and II category structures leads to substantially above-chance responding (optimally, 75% correct for the current structures). This is important because learning is usually taken to mean some relatively permanent adaptation to the environment, rather than simply applying an existing strategy to a novel situation. Hence, one should not use above-chance responding in CJ and II tasks as evidence of category *learning*, because an alternative explanation of above-chance responding is that participants applying an existing strategy. In the current work, the high performance of unidimensional strategies on a conjunction task mean that we may have inadvertently limited the possible performance benefit from training. One possible solution to this issue would be to use category structures where unidimensional strategies score at chance levels (e.g. Donkin et al., 2015).

3.2   Relationship to the COVIS model

According to the COVIS (COmpetition between Verbal and Implicit Systems) model (Ashby, Alfonso-Reese, Turken, & Waldron, 1998), the pattern of near transfer shown in Casale et al. (2012), Zakrzewski et al. (2018) and Soto and Ashby (2019) occurs because different neural systems underlie performance in rule-based (UD and CJ) and II tasks. Participants can analogically transfer knowledge between different rule-based tasks because the Explicit System learns them using rules (Ashby et al., 1998). In contrast, participants fail to transfer knowledge of II tasks because they are learned by the implicit, Procedural System.

The results of the current study neither strongly support nor rule out the COVIS account. On one hand, consistent with COVIS, more people were able to report the correct relationship in the CJ condition than in the II condition (although the Bayesian evidence for this difference was inconclusive). On the other hand, COVIS hypothesizes that the II task is learned implicitly, yet most people reported using a rule-based strategy in this condition (see also Edmunds et al., 2015, 2016, 2019).

Existing published accounts of COVIS are unclear about the possibility of cross-modal transfer of rule-based categories. On the one hand, cross-modal transfer is a form of analogical transfer, and analogical transfer of rule-based categories is expected and has been reported. On the other hand, the specific issue of cross-modal transfer in COVIS has not been covered in either published accounts or formal computational implementations. On the basis of the current

results, we suggest that it may not be necessary to provide COVIS with a mechanism for cross-modal transfer.

3.3 Conclusion

In contrast to previous work on near analogical transfer Casale et al. (2012); Soto and Ashby (2019); Zakrzewski et al. (2018) in perceptual categorization, we found no beneficial effect of far, cross-modal, analogical transfer in either rule-based or information-integration category structures.

3.4 Author contributions

CERE (lead author): Rationale, theory, design, programming, data collection, data analysis, writeup. **ABI**: Rationale, theory, design. Programming and analysis of pilot studies. **PMJ**: Design. **FM**: Rationale, theory, design. Consulted on analysis and write up as co-author. **AJW**: Rationale, theory, design, write up. Consulted on analysis.

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*, 442–81. (https://doi.org/10.1037//0033-295X.105.3.442)

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53. (https://doi.org/10.1037/0278-7393.14.1.33)

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400. (https://doi.org/10.1006/jmps.1993.1023)

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178. (https://doi.org/10.1146/annurev.psych.56.091103.070217)

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147–61. (https://doi.org/10.1111/j.1749-6632.2010.05874.x)

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*(6), 1178–1199. (https://doi.org/10.3758/bf03207622)

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science* (2nd ed., pp. 547–572). New York, NY: Elsevier. (https://doi.org/10.1016/B978-0-08-101107-2.00007-5)

Ashby, F. G., & Valentin, V. V. (2018). The categorization experiment: Experimental design and data analysis. In E. J. Wagenmakers & J. T. Wixted (Eds.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (4th ed., Vol. 5). New York: Wiley. (https://doi.org/10.1002/9781119170174.epcn508)

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*(1), 158–175. (https://doi.org/10.3758/s13428-011-0123-7)

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. (https://doi.org/10.1037/0033-2909.128.4.612)

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. (https://doi.org/10.1038/nrn3475)

Carpenter, K. L., Wills, A. J., Benattayallah, A., & Milton, F. (2016). A Comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*, *37*(10), 3557–3574. (https://doi.org/10.1002/hbm.23259)

Casale, M. B., Roeder, J. L., & Ashby, F. G. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition*, *40*(3), 434–449. (https://doi.org/10.3758/s13421-011-0154-4)

Chase, C. C., Malkiewich, L., & Kumar, A. S. (2019). Learning to notice science concepts in engineering activities and transfer situations. *Science Education*, *103*(2), 440–471. (https://doi.org/10.1002/sce.21496)

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., & Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: a case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 933–48. (https://doi.org/10.1037/xlm0000083)

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Conceptual Structures and Processes: Emergence, Discovery, and Change* (pp. 461–493). Washington D.C.. (https://doi.org/10.1037/10227-017)

Duncker, K. (1945). On problem-solving. (L. S. Lees, Trans.). *Psychological Monographs*, *58*(5), i–113. (https://doi.org/10.1037/h0093599)

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. *The Quarterly Journal of Experimental Psychology*, *68*(2), 1203–1222. (https://doi.org/10.1080/

17470218.2014.978875)

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2018). Due process in dual process: Model-recovery simulations of decision-bound strategy analysis in category learning. *Cognitive Science*, *42*, 833–860. (https://doi.org/10.1111/cogs.12607)

Edmunds, C. E. R., Wills, A. J., & Milton, F. (2016). Memory for exemplars in category learning. In A. Papfragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2243–2248). Austin, TX.

Edmunds, C. E. R., Wills, A. J., & Milton, F. (2019). Initial training with difficult items does not facilitate category learning. *The Quarterly Journal of Experimental Psychology*, *72*(2), 151–167. (https://doi.org/10.1080/17470218.2017.1370477)

Edwards, B. J., Williams, J. J., Gentner, D., & Lombrozo, T. (2019). Explanation recruits comparison in a category-learning task. *Cognition*, *185*, 21–38. (https://doi.org/10.1016/j.cognition.2018.12.011)

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. (https://doi.org/10.3758/bf03193146)

Ford, J. K., Baldwin, T. T., & Prasad, J. (2018). Transfer of training: The known and the unknown. *Annual Review of Organizational Psychology and Organizational Behavior*, *5*, 201–225. (https://doi.org/10.1146/annurev-orgpsych-032117-104443)

Gentner, D., & Maravilla, F. (2018). Analogical reasoning. In L. J. Ball & V. A. Thompson (Eds.), *International Handbook of Thinking & Reasoning* (pp. 186–203). NY, NY.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355. (https://doi.org/10.1016/0010-0285(80)90013-4)

Hornsby, A. N., & Love, B. C. (2014). Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*, *3*(2), 72–76. (https://doi.org/10.1037/e633262013-327)

Johnston, S., Coyer, F., & Nash, R. (2017). Simulation debriefing based on principles of transfer of learning: A pilot study. *Nurse Education in Practice*, *26*, 102–108. (https://doi.org/10.1016/j.nepr.2017.08.002)

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. *Psychology of Learning and Motivation*, *63*, 77–114. (https://doi.org/10.1016/bs.plm.2015.03.001)

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*(1), 49–70. (https://doi.org/10.3758/BF03211715)

Maddox, W. T., Ing, A. D., & Lauritzen, J. S. (2006). Stimulus modality interacts with category structure in perceptual category learning. *Perception & Psychophysics*, *68*(7), 1176–90. (https://doi.org/10.3758/BF03193719)

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*(3), 676–692. (https://doi.org/10.1037/0096-1523.34.3.676)

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs.* Retrieved from https://CRAN.R-project.org/package=BayesFactor (R package version 0.9.12-4.2)

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning: Fact or fantasy? *Psychology of Learning and Motivation*, *54*, 167–215. (https://doi.org/10.1016/B978-0-12-385527-5.00006-1)

R Core Team. (2017). *R: A language and environment for statistical computing.* Retrieved from https://www.R-project.org/

Reetzke, R., Maddox, W. T., & Chandrasekaran, B. (2016). The role of age and executive function in auditory category learning. *Journal of Experimental Child Psychology*, *142*, 48–65. (https://doi.org/10.1016/j.jecp.2015.09.018)

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J. L., Church, B. A., & Ashby, F. G. (2014). Deferred feedback sharply dissociates implicit and explicit category learning. *Psychological Science*, *25*(2), 447–457. (https://doi.org/10.1177%2F0956797613509112)

Smith, J. D., Johnston, J. J. R., Musgrave, R. D., Zakrzewski, A. C., Boomer, J., Church, B. A., & Ashby, F. G. (2014). Cross-modal information integration in category learning. *Attention, Perception, & Psychophysics*, *76*(5), 1473–1484. (https://doi.org/10.3758/s13414-014-0659-6)

Soto, F. A., & Ashby, F. G. (2019). Novel representations that support rule-based categorization are acquired on-the-fly during category learning. *Psychological Research*, *1369*(1), 24–22. (https://doi.org/10.1007/s00426-019-01157-7)

Torchiano, M. (2017). *effsize: Efficient effect size computation.* Retrieved from https://CRAN.R-project.org/package=effsize (R package version 0.7.1)

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. (https://doi.org/10.3758/BF03206482)

Wills, A. J., Inkster, A. B., & Milton, F. (2015). Combination or Differentiation? Two theories of processing

order in classification. *Cognitive Psychology*, *80*, 1–33. (`https://doi.org/10.1016/j.cogpsych.2015.04.002`)

Yi, H.-G., Maddox, W. T., Mumford, J. A., & Chandrasekaran, B. (2016). The Role of Corticostriatal Systems in Speech Category Learning. *Cerebral Cortex*, *26*(4), 1409–1420. (`https://doi.org/10.1093/cercor/bhu236`)

Zakrzewski, A. C., Church, B. A., & Smith, J. D. (2018). The transfer of category knowledge by macaques (Macaca mulatta) and humans (Homo sapiens). *Journal of Comparative Psychology*, *132*(1), 58–74. (`https://doi.org/10.1037/com0000095`)

Appendix

Decision-bound strategy analyses

Most studies in the COVIS canon include analysis of participants' strategies. This is required to patch a logical hole in the design of these experiments (Edmunds et al., 2018). Recall that COVIS predicts the existence of two competing learning systems and that each system is hypothesised to optimally learn particular category structures. However, for any given COVIS experiment, there is no guarantee that participants are using the optimal system for the category structure they learn. Therefore, any dissociation in accuracy may reflect participants using the "wrong" system, rather than reflecting properties of the systems under study. To rule this possibility out, studies investigating the predictions of the COVIS model use a decision-bound model-based analysis to check that the strategies participants use are consistent with the category structure and the system they are supposed to be using.

This decision-bound strategy analysis is informed by a multi-dimensional version of signal detection theory: General Recognition Theory (Ashby & Gott, 1988; Ashby & Maddox, 1993). Broadly, it involves fitting several different strategy models to the data from each participant and seeing which one fits the best. Following Casale et al. (2012), here we fitted three types of strategy: unidimensional, general linear classifier and random. Additionally, we also included a conjunction model as this would be the optimum strategy for the rule-based category structure used here. For ease of exposition, all strategies are framed in terms of the visual stimuli (an example is shown in Figure 2, in the main text).

Unidimensional   This strategy corresponds to a unidimensional rule such as "Stimuli with short lines are in one category, those with long lines are in the other." Two parameters: position of the decision bound and noise.

Conjunction   This strategy corresponds to a conjunction rule such as "Stimuli with short lines high in the box are in Category A, otherwise they are in Category B." Three parameters: two defining the decision boundary plus noise.

General linear classifier   Within the COVIS literature, this strategy is regarded as being learned implicitly. It corresponds to the strategy "If the line is longer than it is high then it is in Category A, otherwise the stimulus is in Category B." Three parameters: two defining the decision boundary plus noise.

Random   Two random models, one with one parameter (bias), one with no parameters.

The best fit of these models was determined using the Bayesian Information Criterion. The best fitting models for each participant are shown in Table 5. Also reported are the Bayesian weights for the success of each model (Wagenmakers & Farrell, 2004). Code for these models fits are available at `https://osf.io/mcd7v`

Notes on interpretation

The results of this strategy analysis are reported here for completeness and for comparison with other work investigating the COVIS model. However, as mentioned briefly in the main text, we have serious concerns about whether this form of strategy analysis accurately reflects the strategies that participants are using. Previous work, by ourselves and others, have shown that the output of this analysis critically depends on the category structure under consideration as well as the strategy models included, regardless of the strategies actually used by the participants (Donkin et al., 2015; Edmunds et al., 2018). For instance, Edmunds et al. (2018) found that approximately 50% of simulated participants who learned an information-integration category structure using a rule-based strategy were misidentified by the analysis as using the optimum, diagonal strategy. Therefore, the strategies reported here should be interpreted very cautiously.

TABLE 5: Decision-bound strategy analysis.

	Block	Strategy *(wBIC)*			
		UD	CJ	GLC	RAND
Conjunction	4	0.10 *(0.70)*	0.77 *(0.84)*	0.13 *(0.60)*	-
	Test	0.57 *(0.72)*	0.13 *(0.64)*	0.07 *(0.47)*	0.23 *(0.72)*
	Control	0.86 *(0.69)*	-	0.03 *(0.54)*	0.10 *(0.79)*
Information-integration	4	0.28 *(0.72)*	0.10 *(0.68)*	0.62 *(0.92)*	-
	Test	0.72 *(0.72)*	0.07 *(0.49)*	0.17 *(0.73)*	0.03 *(0.76)*
	Control	0.97 *(0.72)*	-	-	0.03 *(0.77)*

Strategies: UD=Unidimensional, CJ=Conjunction, GLC=Diagonal general linear classifier, RAND=Random.