Applied Cognitive Psychology, Appl. Cognit. Psychol. **25**: 682–691 (2011)
Published online 5 August 2010 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/acp.1739

Preparing for Novelty with Diverse Training

ANGELA BRUNSTEIN* and CLEOTILDE GONZALEZ

Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Carnegie Mellon University, USA

Summary: This study investigated the ability to generalize acquired skills from training conditions to novel conditions, in a complex perceptual and cognitive task of luggage screening. We examined category and exemplar diversity during training for preparing learners to detect novel items during transfer. Category diversity was manipulated in terms of heterogeneity of training categories: Participants either trained with targets from one category or with targets from several categories. Exemplar diversity was manipulated between participants by presenting either a few or many exemplars for both category diversity conditions. Seventy-two participants were trained to identify threats in pieces of luggage. Thereafter they were transferred to novel stimuli. Results can be summarized in support for the diversity of training hypothesis for preparing for novelty: To the best training for novel luggage screening situations is achieved using fewer items in a variety of categories. Copyright © 2010 John Wiley & Sons, Ltd.

Target detection and decision-making tasks are pervasive. They are as common as finding flaws as a quality inspector, and as important and relevant for our health and society, such as a physician identifying a tumour on an X-ray image, a soldier determining the presence of a combatant in unfamiliar terrain, and an airport security officer looking for threats in passenger luggage.

The terrorist attacks of 11 September changed the way security is addressed in American airports. However, much of the threat detection in luggage screening is still done by visual inspection rather than by automated methods. This is partly due to the complexity of visual images, the uncertainty and variability of what constitutes a threat, and the intricacies of the human decision-making process. Research that improves human accuracy of detecting potential threats and optimizes detection time has become a priority. Our applied research goal in the airport security context is to find ways to transfer skills acquired during training to the accurate detection of unfamiliar, novel targets.

This implies the need to prepare security officers to identify novel items of known categories. For instance, they should be able to detect not only images of guns or knives they have encountered during training but also novel images of guns and knives that they have not yet seen before. We will refer to this as *exemplar diversity* in the following. As we describe below, there exists some evidence that humans can indeed learn to detect novel items of familiar categories. More challenging, security officers have to prepare for another kind of novelty that we call *category diversity*. They have to detect not only novel items of familiar categories, but also novel items of categories that by definition cannot be practiced during training and that will potentially not look like any weapons encountered in previous training. However, we would not expect luggage screeners to be prepared to detect *any* novel object but a novel exemplar of a novel category within a meta-category such as weapons or threats. For example, the meta-category of 'cutting instruments' that are not allowed in an aircraft includes knives, box cutters, machetes, etc. For this category, luggage screeners are likely to be trained on a subset of objects. This subset of objects should help screeners to detect other members of the meta-category of 'cutting instruments'. Similarly, we would expect that when screeners are trained in multiple categories of weapons or threats, they would be able to detect other members of a more generic meta-category of 'threats'. In this research we investigated how to best train to detect this kind of novel items. To do so, we focused on category diversity and exemplar diversity of stimuli presented during training in a luggage screening task to facilitate detection of novel exemplars of a novel category.

There exist several research areas relevant to the goal of preparing luggage screeners for detecting novel items, but none of those provide sufficient and theoretically sound advice to solve this problem. In the following, we will briefly review the literature on skill acquisition, especially for perceptual learning, transfer of skills and category learning, and the implications for preparing for exemplar and category diversity.

Skill acquisition and perceptual learning

When training luggage screeners, we aim to teach skills associated with the task, like visual search, discriminating between targets and distractors, and identifying targets. The skill acquisition literature suggests that with extensive practice, participants improve performance roughly following a power function (Anderson, 1982; Anderson, Fincham, & Douglass, 1999; Logan, 1988; MacKay, 1982; Newell & Rosenbloom, 1981). This effect holds for several tasks, including visual search performance (e.g. Corneille, Goldstone, Queller, & Potter, 2006; Shiffrin & Schneider, 1977; Wolfe, Friedman-Hill, Stewart, & O'Connell, 1992). For example, Gauthier and colleages (Gauthier & Tarr, 1997, 2002; Gauthier, Williams, Tarr, & Tanaka, 1998; Rossion, Gauthier, Goffaux, Tarr, & Crommelinck, 2002) demonstrated that participants could acquire perceptual expertise on a novel category of visual stimuli designed for that series of experiments (Greebles). Most important, Gauthier et al.'s work informs on skill acquisition with systematically manipulated similarity between exemplars. For instance, Gauthier and Tarr (1997) tested performance for transformations within acquired categories but did not test transfer performance for

*Correspondence to: Angela Brunstein, Carnegie Mellon University, Qatar Office, 5032 Forbes Ave SMC1070, Pittsburgh, PA 15289, USA.
E-mail: angelab@cmu.edu

a new category or family of Greebles. More generally, skill acquisition research has focused mainly on ways to improve speed and reduce errors for a given task, rather than on creating flexibility and robustness of knowledge with which humans can address novel task conditions with enhanced performance. For instance, Smith et al. found that participants relied heavily on the recognition of familiar exemplars in visual search and had great difficulty using category-general knowledge (Smith, Redford, Gent, & Washburn, 2005; Smith, Redford, Washburn, & Taglialatela, 2005). As discussed by Chi (2006), expertise, which results from skill acquisition, suffers from many shortcomings including inflexibility of knowledge (*cf.* Singley & Anderson, 1989).

Transfer of skills

Transfer studies extend beyond skill acquisition theories by investigating how acquired skills and knowledge can be applied to novel conditions. For luggage screening, we want officers to transfer visual search skills they acquired with training targets to novel targets they have not seen before. By definition, training and transfer situations have to share some aspects to allow *transfer* of acquired skills and they must be dissimilar to some degree to allow transfer of skills to *novel* conditions. That balance between similar and dissimilar aspects is central to transfer theories. Focusing on the similarity aspect, Healy and colleages (Healy, Wohldmann, Parker, & Bourne, 2005; Healy, Wohldmann, Sutton, & Bourne, 2006) found that individuals displayed retention and transfer of performance for an eye-hand coordination task only when the mental procedures developed during training could be reinstated (i.e. duplicated) at testing. Similarly, Singley and Anderson (1989; see also Sweller, 1980) argue that procedures learned during training have to be shared between training and transfer situations to apply those procedures, for instance, to transfer learned skills to a novel type of text editors.

Given similarity between training and transfer situations, transfer theories agree with skill acquisition theories and their implications for optimal performance. Where both accounts differ is the role of variability in terms of differences between training and transfer situations that defines transfer. Schmidt and Bjork (1992) argued that what works best for improving performance during training does not necessarily coincidence with conditions for optimal performance during transfer. They reviewed evidence for motor and verbal tasks showing that variability during training can enhance performance during transfer or test. For example, variability in the task's order, in the nature and scheduling of feedback often produce better transfer than consistent conditions (Schmidt & Bjork, 1992, see also Schmidt & Lee, 2005). More evidence for the effects of variability during training comes from Doane and colleages (Doane, Sohn, & Schreiber, 1999; Pellegrino, Doane, Fischer, & Alderton, 1991) for learning to discriminate complex polygons. In that case, transfer performance was better for the more difficult discrimination between similar polygons than for the easier discrimination between distinct polygons. Another manipulation increasing the training environment complexity by adding clutter for a luggage screening task also resulted in improved performance during transfer (Fiore, Scielzo, & Jentsch, 2004).

What all these manipulations had in common was that they increased the task difficulty during training. More importantly, they enhanced the processing of training stimuli in a way that matched with the required processing during transfer. This is also known as 'transfer-appropriate processing' (Morris, Bransford, & Franks, 1977; Roediger, Gallo, & Geraci, 2002). For luggage screening, this means keeping constant aspects of the training situations that also apply to transfer situations and varying aspects that do not transfer to allow skill acquisition that is general enough to apply to novel stimuli. Before we can specify training conditions for luggage screening as preparation for novelty, we need to briefly review the literature on category learning.

Category learning

There is evidence that humans are sensitive to exemplar diversity when learning new categories: More-variable categories are harder to learn than less variable categories (e.g. Fried & Holyoak, 1984; Hahn, Bailey, & Elvin, 2005; Homa & Vosburgh, 1976; Peterson, Meagher, Chait, & Gillie, 1973). At the same time, learning variable categories is a better preparation for generalizing to novel members of that category, especially if they are outside of the range of trained category members (Cohen, Nosofky, & Zaki, 2001; Flannagan, Fried, & Holyoak, 1986; Fried & Holyoak, 1984; Hahn et al., 2005; Homa & Vosburgh, 1976; Posner & Keele, 1968; Rips, 1989). For instance, Hahn et al. (2005) manipulated exemplar diversity for one of two perceptual categories of schematic flower images. Category membership for these stimuli was determined by the flowers' head and stem areas. Participants learned to distinguish between pictures of both categories and took a test afterwards with old and novel stimuli that were either very similar to the prototype or dissimilar. In one training condition with low exemplar diversity, flowers presented for the reference category were very similar to the prototype of that category. In the other training condition, the flowers presented were more diverse and dissimilar from the prototype. Training with high exemplar diversity made learning more difficult than training with low exemplar diversity, but higher exemplar diversity during training also increased generalization to novel stimuli outside the range of trained stimuli during test. What is missing in that study in respect to our luggage screening scenario is that participants had to generalize acquired skills to novel stimuli of the same category and not to stimuli of a novel category. Participants were also presented with one stimulus at a time and did not have to search for exemplars of a learned category among distractors.

What remains unknown in the field of category learning is, first, the question of whether diverse training can promote performance for a novel category, for example, other subcategories within a meta-category. For the Hahn et al. (2005) study, would the ability to discriminate between tall and short marguerites also apply to tall and short roses? For luggage screening, would the ability to detect knives also apply to box

cutters? In that case, the effects of exemplar diversity would have to scale up to category diversity. To our knowledge, there exist no studies investigating that kind of transfer. And second, it is not completely clear how results from category learning apply to visual search tasks. So far, visual search and categorization literature are rarely combined in studies (*cf.* Smith, Redford, Gent, et al., 2005; Smith, Redford, Washburn, et al., 2005; Wolfe, Horowitz, van Wert, Kenner, Place, & Kibbi, 2007) because categorization requires identifying targets presented in isolation while visual search requires discriminating targets from simultaneously presented distractors.

Visual search training

There is some evidence that category membership can guide visual search (e.g. Bauer, Jolicoeur, & Cowan, 1996; Daoutis, Pilling, & Davis, 2006; Wolfe et al., 1992), especially for extensively practiced categories like letters or numbers (Brand, 1971; Egeth, Jonides, & Wall, 1972; Gleitman & Jonides, 1976; Ingling, 1972; Jonides & Gleitman, 1972, 1976). Moreover, there is some evidence for the benefits of diverse training for visual search tasks: For a luggage screening scenario, Wolfe et al. (2007) found that searching for high frequency exemplars of a category can promote performance when searching for low frequency exemplars of the same category. This result is especially encouraging because the category used was a meta-category of tools (pliers, axes, drills and hammers) and it seems that benefits in one sub-category transferred to another sub-category in that experiment. However, searching for high frequency exemplars in one category did not help with searching for low frequency exemplars of another unrelated category in another experiment of that series, indicating that there exist limits to diverse training. This pattern makes sense as we would not expect that searching for some targets would help with searching any other targets but only targets from the same or related categories given the evidence we have reviewed thus far.

Smith et al. (Smith, Redford, Gent, et al., 2005; Smith, Redford, Washburn, et al., 2005) investigated a visual search task with dot patterns. Both series of studies are remarkable because the authors successfully combined paradigms from categorization and visual search research to investigate the benefits of diverse training on transfer at the level of exemplar diversity. As with Hahn et al. (2005) in a study on perceptual category learning, high exemplar diversity resulted in worse performance when detecting stimuli during training than low exemplar diversity: Stimuli dissimilar from the prototype of a category (31% correct) were detected less often than stimuli that were similar or identical to the prototype of that category (both 80% correct). For transfer, the authors (Smith, Redford, Washburn, et al., 2005) noted that performance in a learning study dramatically dropped when targets were replaced with similar but new members of a category during the course of the experiment (i.e. introducing exemplar diversity). In this series of studies, participants' performance recovered after the introduction of novel stimuli before dropping again with the introduction of a second set of new targets. Remarkably, participants still performed better with novel targets than baseline in both cases (6–14% improvements of average performance for both blocks with novel stimuli compared to the initial block), indicating there are benefits of diverse training for detecting novel targets of a known category. This was especially true for a luggage screening version of the task instead of searching for dot patterns (14% improvement) or searching for dot-patterns after the location of a potential target was cued before each trial (11% improvement).

In summary, there is some evidence that diverse training as suggested by Schmidt and Bjork (1992) might be beneficial when preparing luggage screeners for detecting novel targets, given that diverse training supports transfer-appropriate processing in terms of preventing too specific skill acquisition. This effect holds true for generalization to novel stimuli within a category (Hahn et al., 2005; Smith, Redford, Gent, et al., 2005; Smith, Redford, Washburn, et al., 2005; Wolfe et al., 2007, experiment 3), but not necessarily for transfer between any categories (Wolfe et al., 2007). It also only holds for tasks where diversity adds to task difficulty, like categorization and visual search, but not for tasks where diversity of stimuli makes the task easier as for same/different judgments (e.g. Ashworth & Dror, 2000; Doane et al., 1999; Pellegrino et al., 1991).

For this study, we extended evidence for category diversity, in addition to exemplar diversity, to prepare learners to detect novel items of a novel category. As Wolfe et al. (2007) demonstrated, it is not very likely that searching for one category like clocks might benefit searching for items of a completely unrelated category like weapons. In our study, we used items that are prohibited in passengers' luggage at the airport. Within that meta-category, we focused on diversity of exemplars during training as well as on similarity between training and transfer situations. We tested the effects of category diversity on performance during training and transfer by introducing three training conditions. We defined one sub-group of threats, tools, as the transfer category. As a *control* condition, one group of participants trained with exemplars from that transfer category right away (low category diversity, same category). A second group also trained also with exemplars from a category that was different from the transfer category (*low category diversity*, different category). To discriminate between these two low category diversity groups, we will label the first group as *control* and the second group as *low category diversity.* And a third group trained with exemplars from several categories that were all different from the transfer category (high category diversity, different categories). In addition, we manipulated exemplar diversity by presenting half of the participants in each condition with a few exemplars (*low exemplar diversity*) and the other half of the participants training with many exemplars for each set of targets (*high exemplar diversity*). We tested the following training hypotheses (see also Table 1):

> *Hypothesis 1*: Searching for items from one category (i.e. control and low category diversity) will result in better training performance than searching for items from several categories (i.e. high category diversity).

Table 1. Hypotheses for training and transfer performance for the three category diversity conditions and the two exemplar diversity conditions

Training performance	Transfer performance
H1 category diversity: control = low > high	H3 category diversity: low < high
	H3a similarity: low < control = high
	H3b heterogeneity: low = control < high
H2 exemplar diversity: low > high	H4 exemplar diversity: low < high

Hypothesis 2: Searching for a restricted number of exemplars per category (i.e. low exemplar diversity) will result in better performance during training than searching for a higher number of exemplars (i.e. high exemplar diversity).

We considered low exemplar diversity during training to be a precondition for successfully learning to perform the task as demonstrated by Smith, Redford, Gent, et al. (2005). At the same time with low exemplar diversity, participants are at risk of acquiring competences specific to trained exemplars that do not transfer well to novel exemplars of that category (Hahn et al., 2005; Smith, Redford, Washburn, et al., 2005). Therefore, as the *diversity of training hypothesis*, we expected high category and exemplar diversity to result in better transfer performance than low category diversity and exemplar diversity. Correspondingly, we tested the following hypotheses:

Hypothesis 3: Searching for items from one category during training (i.e. control and low category diversity) will result in worse transfer performance than searching for items from several categories during training (i.e. high category diversity).

Hypothesis 4: Searching for a restricted number of exemplars per category during training (i.e. low exemplar diversity) will result in worse transfer performance than searching for a higher number of exemplars during training (i.e. high exemplar diversity).

For category diversity, there are two different justifications for the diversity of training hypothesis with one relying on similarity between training and transfer stimuli and the other relying on different kinds of learning outcomes in homogeneous and diverse training participants, where with heterogeneous training fosters more abstract, conceptual learning and homogeneous training fosters more exemplar based, perceptual learning. The *similarity interpretation of the diversity of training hypothesis* suggests that if the application situation is not known in advance, the chances of having encountered something similar before are higher with heterogeneous training that gathers diverse regions of the problem space, than with homogeneous training that gathers just one region of the problem space, except for a special case with novel items either within the range of trained homogeneous stimuli or very close to those. This interpretation is consistent with evidence that transfer performance is best with consistent training with identical conditions for training and transfer (see Healy et al., 2006). Therefore, we tested the following hypotheses:

Hypothesis 3a: Searching for items from several categories (i.e. high category diversity) during training will result in better transfer performance than searching for items from a category that is different from the transfer category (i.e. low category diversity) during training, but will result in worse transfer performance than searching for items from the transfer category right away (i.e. control) during training.

In contrast, the *heterogeneity interpretation of the diversity of training hypothesis* provides a stronger claim by stating that homogeneous and heterogeneous training result in different learning experiences and outcomes. Hahn et al. (2005) demonstrated that identical items during training are not necessary to achieve the best transfer to novel items. In that study, better generalization to novel items of the same category was obtained after heterogeneous training with distant members of that category, compared to homogeneous training that are close to prototype members of that category. For this scenario, one would expect that diverse training with heterogeneous experiences makes people more likely to acquire more general concepts that apply to novel situations than homogeneous experiences. We therefore tested:

Hypothesis 3b: Searching for items from several categories (i.e. high category diversity) during training will result in better transfer performance than searching for items from one category that is either the transfer category (i.e. control) or different from the transfer category (i.e. low category diversity).

METHOD

Participants

Seventy-two college students (47 male, 25 female; $M = 25$ years, $SD = 7$) participated in this study. All participants were right-handed, had normal colour vision, and had normal or corrected-to-normal visual acuity. All participants were recruited from local universities and were paid a total of $15 for their participation. Twelve participants were randomly assigned to one of the six experimental conditions, including a training session and a transfer session taking place 24 hours later. The total participation time did not exceed 1.5 hours.

Design and materials

This study was conducted as a 3×2 between-subjects design with category diversity (control, low, high) and exemplar diversity (low, high) manipulated between the participants during training. Participants were assigned to one of six training conditions and all participants were then transferred to the same transfer condition with novel exemplars that were not shown during the training. Based on a pilot study with

eight participants, 65 objects from three categories were selected as targets for this study. In the pilot, participants arranged pictures of potential targets in a similarity space. Targets for the training conditions were chosen so that items of each category were more similar to each other than to exemplars of other categories and all categories about equally similar to the transfer category of tools.

Correspondingly, the *control* condition involved items drawn from the transfer category of tools (see Figure 1a). The *low category diversity* condition involved items drawn from one category (i.e. knives, see Figure 1b) that was different from the transfer category. The *high category diversity* condition involved objects from four categories (knives, guns, scissors and sharp glass objects; see Figure 1c) that were also different from the transfer category. The *low exemplar diversity* condition involved five different items while the *high exemplar diversity* condition involved 20 different items. Correspondingly, participants trained either with 5 or 20 knives (low category diversity with low *vs*. high exemplar diversity), with 5 or 20 tools (control with low *vs*. high exemplar diversity) or with 5 or 20 objects (high category diversity with low *vs*. high exemplar diversity).

For the transfer session, five novel tools were used as targets. These were novel items that did not appear at all during training.

Our luggage screening simulation is similar to the one used in Wolfe et al. (2007) and Smith et al. (Smith, Redford, Gent, et al., 2005; Smith, Redford, Washburn, et al., 2005). It imitates some of the aspects of the task performed by security officers at the airport when checking passengers' luggage for potential threats. The simulation we developed involves complex visual images of bags (see Figure 2) built from individual X-ray images of targets and distractors. Individual images of these items were provided by the Transportation Security Administration (TSA). The bags images were created using Adobe Photoshop by adding individual objects to images of empty bags and attempting to make them as comparable in clutter as possible (Madhavan & Gonzalez, 2007, 2009). Other studies have used similar luggage screening images to examine the visual search aspects of a screener's performance (i.e. Fiore et al., 2004; McCarley & Carruth, 2004; McCarley, Kramer, Wickens, Vidoni, & Boot, 2004; Washburn, Taglialatela, Rice, & Smith, 2004; Wolfe et al., 2007) for issues related to operator trust in automated decision support systems (Madhavan & Wiegmann, 2005), and categorization and specificity of practice (Smith, Redford, Gent, et al., 2005).

A 'trial' in the luggage screening simulation requires participants to observe a bag filled with various everyday objects (e.g. clothes, hair dryers and pill bottles). In 50% of all cases the bags contained one target, such as a knife, gun, scissors, sharp glass object or tool. Participants had to inspect these bags and click on detected targets or let the presentation of the bag time out after 4 seconds. Immediately after each trial, the system provides participants with feedback in form of a text message saying 'weapon identified' (hit), 'no weapon at location' (false alarm), 'weapon missed' (miss) or 'secure bag cleared' (correct rejection). Note that a false alarm message is displayed for both clicking a distractor in a secure bag and clicking in a bag with a target in another location. Next to the text message, there is the score per trial displayed with '10' for correct responses and '−10' for incorrect responses, and the accumulated score for the current block of 100 trials. There was no other form of feedback implemented. Participants were informed whether or not the bag contained a target, but not which item was the target. The bag was not shown again after failure and there is no chance for participants to correct their response.

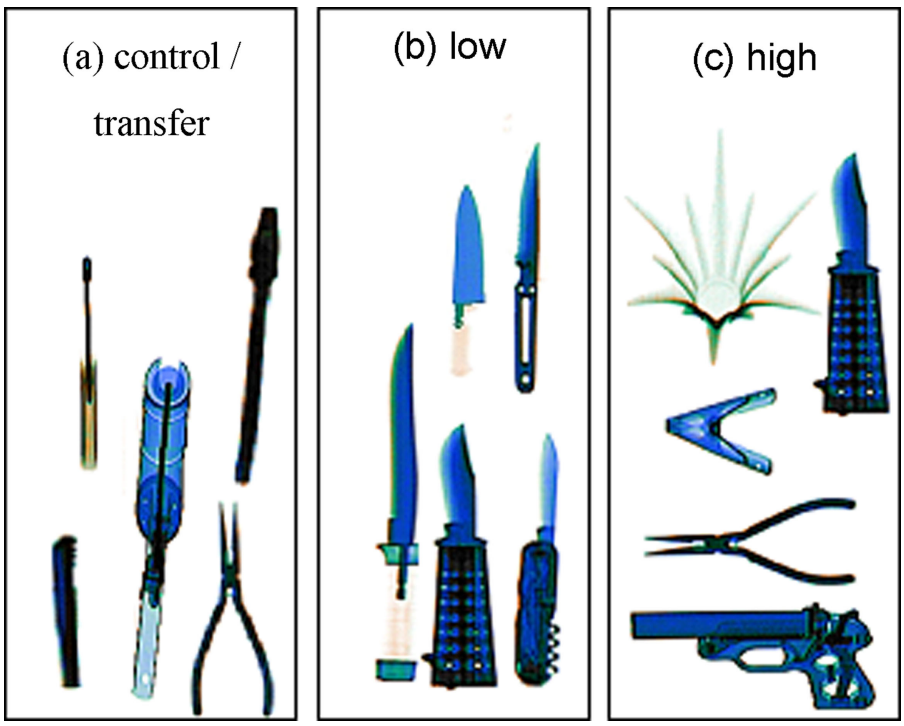

Figure 1. Examples of targets used in the luggage screening task. In the training session of experiment 1, participants were assigned randomly to one of three category diversity conditions: (a) control, (b) low or (c) high. At transfer, objects from set (a) were used as targets

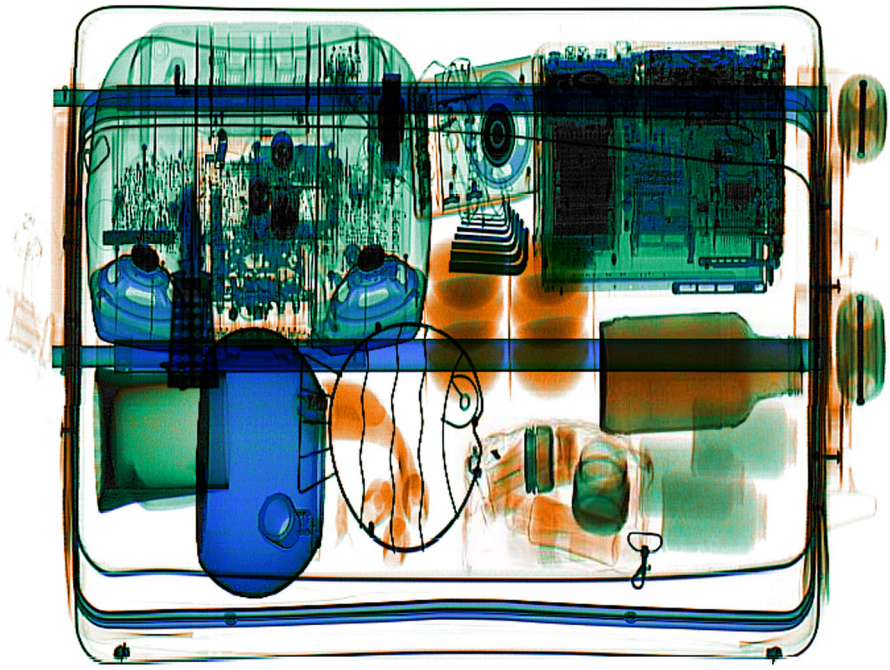

Figure 2. A bag used for the luggage screening task, composed from individual X-ray images provided by the TSA. This bag contains the image of a knife in the upper left corner and is therefore a target bag

Procedure

The training session lasted about 1 hour and consisted of four blocks with 100 trials each and a base rate of 50% for targets. For training with low exemplar diversity (i.e. training with five knives, five tools or five objects), all four blocks had the same memory set of five items to search for. For training with high exemplar diversity (i.e. training with 20 knives, 20 tools or 20 objects), each block had a different memory set of five items to search for. During the training session, participants were asked to memorize a set of five targets before each of four blocks. Then, they were asked to search for any member of that set that appeared in the luggage for 100 trials.

The transfer session lasted half an hour and was run 24 hours after training. The transfer session consisted of one block with 100 trials and also a base rate of 50% for targets. During the transfer session, participants were not shown a memory set of targets to search for. Instead, they were instructed to use their best judgment to detect 'any possible dangerous items in the bags'. For all participants, that definition of targets was accepted as sufficient to perform the task. In order to keep conditions as similar as possible between training and transfer, participants received textual feedback about the accuracy of their diagnosis immediately after the end of each trial but were not shown the targets in case of failure nor were they given a chance to correct their responses.

Dependent variables

The dependent variable was detection accuracy, measured by hit rates and false alarm rates. Based on these rates we calculated d' and beta to measure overall performance. In accordance categorization research (e.g. Smith, Redford, Gent, et al., 2005; Smith, Redford, Washburn, et al., 2005), we counted each click on a target as a correct response. Using this criterion, we might have missed a couple of identified targets with slightly misplaced clicks. We tried to accommodate for that by defining the target location a little bigger than the exact target area.

For the training session, we expected participants in the low category and control conditions to perform better in terms of higher d' than participants in the high category condition (see Hypothesis 1). In addition, participants in the low exemplar diversity condition should perform better than participants in the high exemplar diversity condition (see Hypothesis 2). More importantly for the transfer session, according to our diversity of training hypothesis (see Hypothesis 3), we expected participants in the high category diversity condition to perform better in terms of higher d' than participants in the low category condition when searching for novel targets. Participants in the control condition should either perform best during transfer (see Hypothesis 3a) or as poorly as participants in the low category diversity condition (see Hypothesis 3b). When detecting novel targets, participants in the high exemplar diversity condition should also perform better than participants in the low exemplar diversity condition for detecting and identifying targets (see Hypothesis 4).

RESULTS

For training performance, we conducted ANOVAs with blocks (4) as repeated measures factor and category diversity (control, low, high) and exemplar diversity (low, high) as between-subjects factors for d' and beta for identifying targets. For transfer performance we conducted ANOVAs with category diversity (3) and exemplar diversity (2) as

between-subjects factors for d' and beta for identifying targets. For *post hoc* comparisons, we used Bonferroni tests.

Training

Table 2 shows the performance of identifying targets during training for each of the six conditions.

d' for identifying targets during training

The analysis of variance for d' revealed a significant main effect of blocks on training performance for identifying targets, $F(3, 198) = 43.14$, $p < .01$, $\eta^2 = .40$. This implies participants improved their performance with practice.

In addition, the analysis showed main effects of category diversity, $F(2, 66) = 9.45$, $p < .01$, $\eta^2 = .22$, and of exemplar diversity on training performance, $F(1, 66) = 151.85$, $p < .01$, $\eta^2 = .70$, and an interaction between both, $F(2, 66) = 9.90$, $p < .01$, $\eta^2 = .23$. Participants in the low category diversity condition performed as well during training as participants in the high category diversity condition ($p = 1.00$) and both performed better (p's $< .01$) than control. It seems that the category of tools was harder to acquire than both knives and several other categories of objects. However, all participants in all groups improved their performance during training demonstrating significant learning. Participants in the low exemplar diversity condition performed better during training than participants in the high exemplar diversity condition (see Table 2).

Beta for indentifying targets during training

The variance analysis for beta showed no effect of blocks on training performance, $F(3, 198) = 0.97$, $\eta^2 = .01$, indicating that participants' criterion for correctly identified targets remained stable across training blocks.

In addition, the analysis showed main effects of category diversity, $F(2, 66) = 4.45$, $p < .05$, $\eta^2 = .12$, and of exemplar diversity on training performance, $F(1, 66) = 4.99$, $p < .05$, $\eta^2 = .07$, and an interaction between both, $F(2, 66) = 3{,}26$, $p <= .05$, $\eta^2 = .09$. The main effect of category diversity indicates that participants in the low category diversity conditions demonstrated a more conservative criterion for reporting targets ($\beta = .89$) than participants in the high category diversity conditions ($\beta = .44$). Participants in the low exemplar diversity conditions demonstrated a more liberal criterion ($\beta = .55$) than participants in the high exemplar diversity conditions ($\beta = .82$).

Summary training

Participants improved their performance, as measured by d', during training. Against Hypothesis 1, control participants performed worse than participants in low and high category diversity conditions. That was especially true for control participants in the high exemplar diversity condition. It seems that tools as a category were harder to identify than knives or several objects. The analysis for beta revealed that performance of participants in the low category diversity conditions did not differ from participants in the control conditions, but participants in the low category diversity conditions demonstrated a more conservative criterion for identifying targets than participants in the low category diversity conditions. Therefore, Hypothesis 1 holds true for low category diversity producing better training performance than high category diversity, but not for control. Regarding Hypothesis 2, low exemplar diversity resulted in better identification rates than high exemplar diversity as expected. We interpret this as minimum criterion for sufficient learning.

Transfer

Table 3 shows performance of the six groups during transfer.

d' for identifying targets during transfer

The analysis of variance for d' revealed significant main effects of category diversity, $F(2, 66) = 5.15$, $p < .01$, $\eta^2 = .14$, and of exemplar diversity on transfer performance for identifying targets, $F(1, 66) = 8.60$, $p < .01$, $\eta^2 = .12$, but no interaction between both, $F(2, 66) = 0.83$, $\eta^2 = .02$. This implies participants in the low category diversity condition performed as well as control participants ($p = 1.00$) and both performed worse (ps $< .05$) than participants in the high category diversity condition. Again, participants in the low exemplar diversity condition performed better when

Table 2. Performance during the training session in terms of d' and response times

Dependent variable	Condition		Performance mean (SE)
	Category	Exemplar	
d'	Low	Low	2.96 (.24)
		High	1.63 (.19)
	Control	Low	2.99 (.25)
		High	.45 (.24)
	High	Low	2.99 (.23)
		High	1.81 (.29)
Beta	Low	Low	0.97 (.40)
		High	0.80 (.11)
	Control	Low	.46 (.11)
		High	1.00 (.13)
	High	Low	0.21 (.06)
		High	0.67 (.20)

Table 3. Performance during the transfer session

Dependent variable	Condition		Performance mean (SE)
	Category	Exemplar	
d'	Low	Low	1.00 (.34)
		High	0.30 (.37)
	Control	Low	0.65 (.28)
		High	0.29 (.23)
	High	Low	1.98 (.35)
		High	0.83 (.26)
Beta	Low	Low	0.97 (.14)
		High	1.54 (.30)
	Control	Low	1.01 (.13)
		High	1.11 (.10)
	High	Low	0.50 (.07)
		High	1.16 (.36)

identifying novel targets than participants in the high exemplar diversity condition (see Table 3).

Beta for identifying targets during transfer

A variance analysis for beta revealed no effect of category diversity, $F(2, 66) = 1.99$, $p = .14$, $\eta^2 = .06$, but a main effect of exemplar diversity on transfer performance, $F(1, 66) = 6.60$, $p < .05$, $\eta^2 = .09$, and no interaction between both, $F(2, 66) = 0.98$, $\eta^2 = .03$. This implies participants in the low exemplar diversity condition ($\beta = .83$) demonstrated a more conservative criterion for identifying targets than participants in the high exemplar diversity condition ($\beta = 1.27$).

Summary transfer

As expected by Hypothesis 3, participants in the high category diversity condition performed best during transfer, according to d'. In contrast with Hypothesis 3a, participants in the low category diversity condition performed as well during transfer as control participants. This confirms the heterogeneity version of the diversity of training Hypothesis 3b: Better performance by participants in the diverse training conditions is not primarily based on similarity of some training experiences with transfer experiences, but on the heterogeneity of experiences during training to prevent overly specific exemplar learning. It is remarkable that participants in the control conditions performed worst during transfer, despite being trained on exemplars of the transfer category all along. This is probably due to the task difficulty of identifying tools, compared to detecting knives or diverse objects. Against Hypothesis 4, participants in the low exemplar diversity conditions performed better than participants in the high exemplar diversity condition not only during test but also during transfer. With respect to response criteria, there was an effect of exemplar diversity during training on the criterion demonstrated for identifying targets. The betas changed dramatically from more liberal during training to more conservative during transfer. This makes sense because participants were told what to look for during training but not during transfer.

DISCUSSION

This study investigated the effects of category and exemplar diversity during training on performance during transfer for novel image recognition situations. For example, for airport luggage screeners, these novel situations are novel threats, for which by definition, training cannot be carried out in advance. More globally, for most training scenarios, learners have to generalize from experiences during training to novel situations outside the classroom.

Based on the results of our study, we conclude that employing a few and not many items from diverse categories and from more than one homogeneous category during training will best prepare luggage screeners for these novel threats. It appears that better recognition of novel items comes from recognizing fewer items encountered during training than from recognizing a broader range of less well-trained items contradicting Hypothesis 4. This result seems to also contradict with evidence for the benefits of exemplar diversity presented by Hahn et al. (2005) and Smith, Redford, Washburn, et al. (2005). However, our manipulation of exemplar diversity is slightly different from both of the previous investigations: Hahn et al. and Smith et al. did not manipulate the number of exemplars but their distance from the prototype of a category, thus paralleling within a category what we manipulated between categories. They found that the more dissimilar from the prototype exemplars, the more learning was impaired. This matches our results for category diversity, but not our results for exemplar diversity.

Secondly and more importantly, it is better to train with diverse categories and not just on one category to prepare for novelty confirming Hypothesis 3. This effect of category diversity cannot be explained just by the similarity between training and transfer targets (see Hypothesis 3a), but it needs to be explained by the heterogeneity of experience during training (see Hypothesis 3b). This result provides strong support for our diversity of training hypothesis. In terms of categorization theories, diverse training can hinder overly specific exemplar learning. This result confirms Hahn et al.'s (2005) finding that participants better generalized to novel exemplars after diverse training than after homogeneous training within a category. It also matches with Wolfe et al.'s (2007) finding benefits for diverse training within a meta-category but not between unrelated categories.

One interesting aspect of performance concerns adaptation in terms of sensitivity and response criterion when facing novel stimuli. While Wolfe et al.'s (2007) experiment 3 implies that participants changed their response criterion for different objects within a category, Hahn et al. (2005) found that participants in the diverse training condition differed from participants in the homogeneous training condition in terms of sensitivity, but not in terms of their response criterion. In our study we found a different pattern for category *versus* exemplar diversity conditions. We found similar betas for low *versus* high category and exemplar diversity conditions. For category diversity conditions, we found the same beta, but higher d's for high than low category diversity conditions. In contrast, for exemplar diversity conditions, we found both higher betas and higher d's for low than for high exemplar diversity conditions. It seems that participants tried their best to cope with the novel and difficult conditions during transfer. Participants who were worse prepared for novel stimuli based on their sensitivity seem to have adjusted their response criterion accordingly. More research will be needed to explore that effect more carefully and to fully understand this phenomenon.

Another interesting aspect concerns transfer. Our data would not support generalization to novel targets by adding any objects. Wolfe et al. (2007) did not find consistently better performance for detecting rare weapons by adding unrelated items, but did find better performance only by adding other kinds of tools to a rare sub-category of tools. For our study, adding more exemplars of diverse objects or of tools did not improve either training or transfer performance, probably because what it takes to identify threats are training targets that highlight critical attributes. Because learning diverse objects takes more effort than homogenous

categories, the number of items used should be as small as possible but representing all critical attributes. On the other hand, repeated exposure to the attributes in several objects helps to separate critical from non-critical features and this could result in improved performance.

In addition, the category of tools was the hardest to acquire during training and still difficult during transfer even for participants who had trained with that category all along. This might sound surprising because tools should be objects most participants are familiar with. In this study, however, students were presented with X-rays of tools and not actual pictures of tools. In addition, tools form a meta-category with individual items that were hard to name for participants of a pilot study while knives form a basic level category. For knives, those participants happily referred to 'another knife', while they would never refer to 'another tool'. For several objects, participants might have memorized something like 'two knives, one gun, a screw driver and a hammer'. For tools it might have been 'a screw driver, a hammer and two things I have never seen before'. It would be interesting to replicate this study with another transfer category that has comparable learning outcomes for all involved categories before participants are transferred to novel stimuli.

In the domain of luggage screening it is very likely that novel targets might be as hard to detect as our tool condition. In related studies on time pressure and workload, Gonzalez (2004, 2005) has found that what best prepares for difficult tasks are not those difficult conditions but conditions that optimally support learning as a specific version of transfer-appropriate processing. Corresponding to those studies it is better to transfer skills acquired under easier conditions to harder transfer conditions than not sufficiently acquiring that skills when training under hard conditions right away.

In summary, the best training for novel luggage screening targets is a restricted number of items collected from a variety of categories, enabling successful learning and preventing over-specialization. There are two limitations to these conclusions. First, what has to be considered as small or large numbers of exemplars critically depends on the amount of training learners are exposed to. For our participants training for about one hour, five items resulted in better performance in the learning session than 20. For luggage screeners in the UK who have had trained for several months, 250 items are already a small enough number resulting in over-specialization (see Smith, Redford, Washburn, et al., 2005, p. 1172) and hindering generalization to novel targets. In that study, threat images were occasionally digitally injected into the images of actual bags passing through security. When introducing novel images of weapons after several months, performance of the luggage screeners dropped back to baseline despite extensive training.

Secondly, when considering category diversity of training, we are at risk of comparing apples to oranges. Even for categories with the same within-category similarity between exemplars, like knives and tools, categories can still differ remarkably and can create very different learning conditions. In our study, training on knives resulted in much better learning outcomes than training on tools. Both sets of targets resulted in comparably poor performance when searching for novel targets. As described above, the task difficulty might interact with the benefits of diverse training. If searching for knives had have been the transfer task we might have gotten different results in this study. This issue could be especially relevant for categories that are created dynamically on demand and are not predefined. Further investigations will be needed to identify all the factors relevant to category diversity affecting preparation for novel situations.

This research also brings back the spotlight on issues discussed by Schmidt and Bjork (1992) concerning existing trade-offs in training speed, accuracy and real-world performance. Training programs must not only test the practice effects of the variables of interest, but also the transfer, durability and generalization of the knowledge acquired when those variables are removed or modified. According to our results, the transfer of training to real-world performance depends directly on the diversity of conditions used in training and on the similarity between training and transfer conditions.

ACKNOWLEDGEMENTS

This research was partially supported by the Multidisciplinary University Research Initiative Program (MURI; N00014-01-1-0677) and by the National Science Foundation (Human and Social Dynamics: Decision, Risk, and Uncertainty, Award number: 0624228) awards to Cleotilde Gonzalez. The authors are also grateful for editorial assistance during the preparation of this paper provided by Lisa Czlonka and programming assistance for the Luggage Screening task provided by Varun Dutt. They thank Frank Ritter for helpful comments on an earlier version of the paper.

REFERENCES

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*, 369–403.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1120–1136.

Ashworth, A. R. S., & Dror, I. E. (2000). Object identification as a function of discriminability and learning presentations: The effect of stimulus similarity and canonical frame alignment on aircraft identification. *Journal of Experimental Psychology: Applied*, *6*, 148–157.

Bauer, B., Jolicoeur, P., & Cowan, W. B. (1996). Distractor heterogeneity versus linear separability in color search. *Perception*, *25*, 1281–1293.

Brand, J. (1971). Classification without identification in visual search. *Quarterly Journal of Experimental Psychology*, *23*, 178–186.

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K. A. Ericson, N. Charness, & P. J. Feltovitch (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 121–130). Cambridge, UK: Cambridge University Press.

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory & Cognition*, *29*, 1165–1175.

Corneille, O., Goldstone, R. L., Queller, S., & Potter, T. (2006). Asymmetries in categorization, perceptual discrimination, and visual search for reference and nonreference exemplars. *Memory & Cognition*, *34*, 556–567.

Daoutis, C. A., Pilling, M., & Davies, I. R. L. (2006). Categorical effects in visual search for colour. *Visual Cognition*, *14*, 217–240.

Doane, S. M., Sohn, Y. W., & Schreiber, B. (1999). The Role of Processing Strategies in the Acquisition and Transfer of a Cognitive Skill. *Journal of*

Experimental Psychology: Human Perception and Performance, *25*, 1390–1410.

Egeth, H., Jonides, J., & Wall, S. (1972). Parallel processing of multi-element displays. *Cognitive Psychology*, *3*, 674–698.

Fiore, S. M., Scielzo, S., & Jentsch, F. (2004). Stimulus competition during preceptual learning: Training and aptitude consideration in the x-ray security screening process. *Cognitive Technology*, *9*, 34–39.

Flannagan, M. J., Fried, L. S., & Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *12*, 241–256.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *10*, 234–257.

Gauthier, I., & Tarr, M. J. (1997). Becoming a 'Greeble' expert: Exploring mechanisms for face recognition. *Vision Research*, *37*, 1673–1682.

Gauthier, I., & Tarr, M. J. (2002). Unraveling mechanisms for expert object recognition: Bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 431–446.

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training 'Greeble' experts: A framework for studying expert object recognition processes. *Vision Research*, *38*, 2401–2428.

Gleitman, H., & Jonides, J. (1976). The cost of categorization in visual search: Incomplete processing of targets and field items. *Perception & Psychophysics*, *20*, 281–288.

Gonzalez, C. (2004). Learning to make decisions in dynamic environments: Effects of time constraints and cognitive abilities. *Human Factors*, *46*, 449–460.

Gonzalez, C. (2005). Task workload and cognitive abilities in dynamic decision making. *Human Factors*, *47*, 92–101.

Healy, A. F., Wohldmann, E. L., Parker, J. T., & Bourne, L. E. Jr., (2005). Skill training, retention, and transfer: The effects of a concurrent secondary task. *Memory & Cognition*, *33*, 1457–1471.

Healy, A. F., Wohldmann, E. L., Sutton, E. M., & Bourne, L. E. Jr., (2006). Specificity effects in training and transfer of speeded responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 534–546.

Hahn, U., Bailey, T.M., & Elvin, L.B.C. (2005). Effects of category diversity on learning, memory, and generalization. *Memory & Cognition*, *33*, 289–302.

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning & Memory*, *2*, 322–330.

Ingling, N. W. (1972). Categorization: A mechanism for rapid information processing. *Journal of Experimental Psychology*, *94*, 239–243.

Jonides, J., & Gleitman, H. (1972). A conceptual category effect in visual search: O as a letter or a digit. *Perception & Psychophysics*, *12*, 456–460.

Jonides, J., & Gleitman, H. (1976). The benefits of categorization in visual search: Target location without identification. *Perception & Psychophysics*, *20*, 289–298.

Logan, G. (1988). Toward an Instance Theory of Automatization. *Psychological Review*, *95*, 492–527.

MacKay, D. G. (1982). The problem of flexibility, fluency, and speed-acuracy trade-off in skilled behavior. *Psyschological Review*, *89*, 483–506.

Madhavan, P., & Gonzalez, C. (2007). Differential base rate training influences detection of novel targets in a complex visual inspection task. In *Proceedings of the human factors and ergonomics society 51st annual meeting* (pp. 392–396). Baltimore, MD: Human Factors and Ergonomics Society.

Madhavan, P., & Gonzalez, C. (2009). The relationship between stimulus-response mappings and the detection of novel stimuli in a simulated luggage screening task. Conditionally accepted for *Theoretical Issues in Ergonomics Science*, 1464–536X.

Madhavan, P., & Wiegmann, D. A. (2005). Cognitive anchoring on self-generated decisions reduces operator reliance on automated diagnostic aids. *Human Factors*, *47*, 332–341.

McCarley, J. S., & Carruth, D. W. (2004). Oculomotor scanning and target recognition luggage x-ray screening. *Cognitive Technology*, *9*, 26–29.

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. *American Psychological Society*, *15*, 302–306.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–56). Hillsdale, NJ: Erlbaum.

Pellegrino, J. W., Doane, S. M., Fischer, S. C., & Alderton, D. (1991). Stimulus complexity effects in visual comparisons: The effects of practice and learning context. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 781–791.

Peterson, M. J., Meagher, R. B., Jr., Chait, H., & Gillie, S. (1973). The abstraction and generalization of dot patterns. *Cognitive Psychology*, *4*, 378–398.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou, & A. Ortney (Eds.), *Similarity and analogical reasoning* (pp. 21–59). New York: Cambridge University Press.

Roediger, H. L., Gallo, D. A., & Geraci, L. (2002). Processing approaches to cognition: The impetus from the levels-of-processing framework. *Memory*, *10*, 319–332.

Rossion, B., Gauthier, I., Goffaux, V., Tarr, M. J., & Crommelinck, M. (2002). Expertise training with novel objects leads to left lateralized face-like electrophysiological responses. *Psychological Science*, *13*, 250–257.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.

Schmidt, R. A., & Lee, T. D. (2005). *Motor control and learning: A behavioral emphasis*. Champaign, IL: Human Kinetics.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: Perceptual learning, automatic attending, and a general theory. *Psychological Review*, *84*, 127–190.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Smith, J. D., Redford, J. S., Gent, L. C., & Washburn, D. A. (2005). Visual search and the collapse of categorization. *Journal of Experimental Psychology: General*, *134*, 443–460.

Smith, J. D., Redford, J. S., Washburn, D. A., & Taglialatela, L. A. (2005). Specific-token effects in screening tasks: Possible implications for aviation security. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *31*, 1171–1185.

Sweller, J. (1980). Transfer effects in a problem solving context. *Quarterly Journal of Experimental Psychology*, *32*, 233–239.

Washburn, D. A., Taglialatela, L. A., Rice, P. R., & Smith, J. D. (2004). Individual differences in sustained attention and threat detection. *Cognitive Technology*, *9*, 30–33.

Wolfe, J. M., Friedman-Hill, S. R., Stewart, M. I., & O'Connell, K. M. (1992). The role of categorization in visual search for orientation. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 34–49.

Wolfe, J. M., Horowitz, T. S., van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General*, *136*, 623–638.