Attention, Perception, & Psychophysics (2019) 81:137–157
https://doi.org/10.3758/s13414-018-1595-7

Semisupervised category learning facilitates the development of automaticity

Katleen Vandist[1,2] · Gert Storms[2] · Eva Van den Bussche[1,2]

Published online: 21 September 2018

Abstract

In the human category of learning, learning is studied in a supervised, an unsupervised, or a semisupervised way. The rare human semisupervised category of learning studies all focus on early learning. However, the impact of the semisupervised category learning late in learning, when automaticity develops, is unknown. Therefore, in Experiment 1, all participants were first trained on the information-integration category structure for 2 days until they reached an expert level. Afterwards, half of the participants learned in a supervised way and the other half in a semisupervised way over two successive days. Both groups received an equal number of feedback trials. Finally, all participants took part in a test day where they were asked to respond as quickly as possible. Participants were significantly faster on this test in the semisupervised group than in the supervised group. This difference was not found on day 2, implying that the no-feedback trials in the semisupervised condition facilitated automaticity. Experiment 2 was designed to test whether the higher number of trials in the semisupervised condition of Experiment 1 caused the faster response times. We therefore created an almost supervised condition where participants almost always received feedback (95%) and an almost unsupervised condition where participants almost never received feedback (5%). All conditions now contained an equal number of trials to the semisupervised condition of Experiment 1. The results show that receiving feedback almost always or almost never led to slower response times than the semisupervised condition of Experiment 1. This confirms the advantage of semisupervised learning late in learning.

Keywords Categorization · Semisupervised learning · Automaticity

Introduction

Throughout the ages correct categorization has remained essential for human survival. In prehistory, classifying a predator as a harmless animal was mortal. Nowadays, classifying traffic signs correctly is important to avoid accidents. These examples explain why categorization has received continuous attention in the field of cognitive science (e.g., Ashby & Maddox, 2005; Ashby & Maddox, 2010; Medin & Schaffer, 1978; Nosofsky, 1987; Pothos & Chater, 2002). Although many categories that people use are acquired during childhood (French, Mareschal, Mermillod, & Quinn, 2004), adults also learn new categories. In the human category of learning research the focus is on the learning process itself. In order to understand this learning process, exemplars and non-exemplars of unfamiliar categories are typically presented (Ashby & Maddox, 2005). The behavior of participants is observed during the period when their ability to assign stimuli to these categories increases from chance level to a certain stable above-chance level (Ashby & Maddox, 2005).

Supervised and unsupervised learning

In the past, the human category of learning was studied using supervised or unsupervised learning paradigms. In *supervised learning paradigms*, the participant is presented with a stimulus that has to be classified into two or more contrasting categories. Immediately after this response, feedback is always provided about the correct category label. Generally the participant knows the number of contrasting categories in advance (see Shepard, Hovland, & Jenkins, 1961 for a

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13414-018-1595-7) contains supplementary material, which is available to authorized users.

✉ Katleen Vandist
katleen.vandist@vub.be

[1] Department of Psychology, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

[2] Department of Experimental Psychology, KU Leuven, Leuven, Belgium

description of a basic experiment). Numerous studies have demonstrated that, based on this paradigm, participants can learn very complex categories if a sufficient number of trials is provided (e.g., Ashby, Queller, & Berretty, 1999; Ashby, Maddox, & Bohil, 2002; Maddox, Filoteo, Hejl, & Ing, 2004c; McKinley & Nosofsky, 1995; Medin & Schwanenflugel, 1981; Maddox, Ashby, & Gottlob, 1998). In *unsupervised learning paradigms*, the participant never receives feedback or information about the category to which the presented stimulus belongs. The goal is to identify an intuitive or natural classification for a set of objects (Clapper & Bower, 1994; Love, 2002; Medin, Wattenmaker, & Hampson, 1987; Milton, Longmore & Wills, 2008; Pothos & Chater, 2002, 2005; Pothos et al., 2011). Depending on the paradigm, the number of contrasting categories may or may not be known in advance. Findings based on such unsupervised paradigms reveal that performance is dominated by the use of unidimensional rules, regardless of the complexity of the underlying category structure or the number of training trials (Ashby et al., 1999). These unidimensional rules (e.g., "small stimuli belong to category A and large stimuli to category B") are easy to verbalize and to apply, whereas complex categorization rules are mostly hard to express. In conclusion, in unsupervised learning people have the tendency to use very simple categorization rules, whereas in supervised learning participants are able to learn very complex categorization structures.

Semisupervised learning

Vandist, De Schryver, and Rosseel (2009) argued that both supervised and unsupervised learning are ecologically rare. Translated to daily life, supervised learning means that for every object that we observe we immediately receive correct information about its category label. In most category learning situations, it seems very unlikely that this occurs after each single encounter of a category member. Strictly speaking, supervised learning would imply that, when we walk in the woods, a label "tree" is attached to every single tree. Moreover, this information is unambiguous, implying that the information provider and receiver always mean the same object. However, ambiguity often occurs in very rich environments. For example, when walking in the woods a parent might point to a bird and call it "bird", while the child may be watching a nest and hence learns the wrong label. Unsupervised learning on the other hand entails that we never receive any information about object categories. This implies that during our entire lives, nobody ever informs us about the name of an object or about which objects belong together. Both types of category learning therefore do not represent our daily reality. In a previous study Vandist et al. (2009) argued that people instead learn in a *semisupervised way*: when confronted with (new) objects (e.g., a dog), sometimes category information will be provided ("look, a dog") and sometimes not. This idea is supported by Gibson, Rogers, and Zhu (2013). In the *semisupervised category learning paradigm* this realistic scenario is incorporated. In a block of trials a predetermined percentage of category responses is followed by feedback (i.e., feedback or labelled trials). The remaining trials do not receive feedback (i.e., no-feedback or unlabelled trials). For example, a block in a 25% semisupervised classification learning paradigm consists of 25% feedback trials and 75% no-feedback trials.

Vandist et al. (2009) compared the effects of this semisupervised learning process to supervised and unsupervised learning processes by using the information-integration structure. This category structure is frequently used in category learning research (e.g., Ashby et al., 2002; Ashby & Ell, 2001; Ell & Ashby, 2006; Maddox, Ashby, Ing, & Pickering, 2004a; Maddox & Filoteo, 2011; Maddox & Ing, 2005; Maddox, Pacheco, Reeves, Zhu, & Schnyer, 2010b; Paul, Boomer, Smith, & Ashby, 2011; Spiering & Ashby, 2008a, b; Vermaercke, Cop, Willems, D'Hooge, & Op de Beeck, 2014). Figure 1 shows an example of the information-integration category structure used in the experiments reported in the current study. Although the within-category correlation is very high, this structure is difficult to learn. To obtain high-level performance, participants have to combine the perceptual information of the underlying stimulus dimensions simultaneously at some predecisional stage (Ashby & Gott, 1988). This perceptual integration could take many forms – in this case, by calculating a weighted linear combination of the dimensional values. The optimal decision bound is almost impossible to describe verbally (Ashby, Alfonso-Reese, Turken,

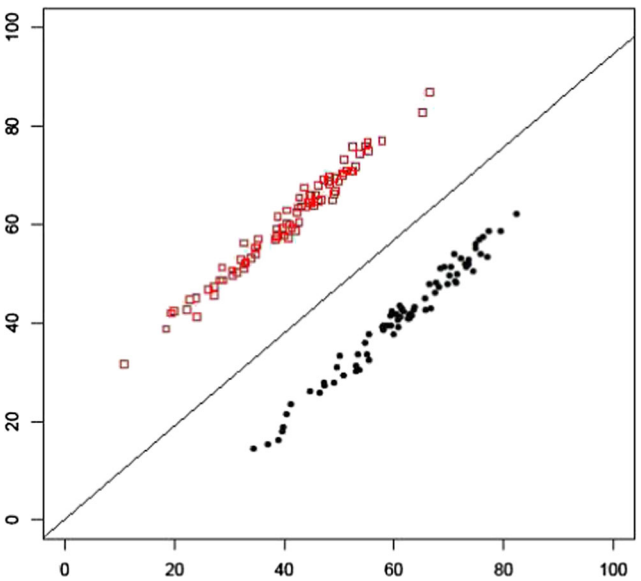

Fig. 1 An example of the information-integration category structure used in the learning task in Experiment 1. The "A" stimuli are shown in squares, the "B" stimuli in solid circles. The decision bound that divides the two categories perfectly is shown in black. The Y-axis is the Orientation dimension, the X-axis the Frequency dimension

& Waldron, 1998) and it cannot readily be discovered via an explicit reasoning process (Ashby & O'Brien, 2007), which makes the category structure difficult to master for humans (Vermaercke et al., 2014). Feedback is essential to learn the structure successfully (Ashby et al., 1999).

The results of the study of Vandist et al. (2009) indicated that, as expected, learning the information-integration structure was successful in the supervised condition but not in the unsupervised condition. In a 50% semisupervised condition, participants managed to learn the structure, suggesting that feedback after every trial is not necessary to learn the complex structure.

The impact of no-feedback trials in semisupervised learning

An additional goal of the study of Vandist et al. (2009) was to understand the contribution of the no-feedback trials on the learning process. In the no-feedback trials, participants were shown a stimulus, processed it, and categorized it. It was only clear after the categorization that no feedback followed. Crucially, it was investigated whether the processing of the stimulus in the no-feedback trials had an impact on the learning process or whether the experience was simply neglected. To achieve this, the no-feedback trials were replaced by irrelevant fillers, where no categorization whatsoever takes place. The number of feedback trials was identical in the condition with no-feedback trials and the condition with filler trials. The results indicated that the learning process in both conditions was similar. Hence, the no-feedback trials neither harmed nor helped learning. Apparently, when we encounter an object early in the learning process, we classify it, and when no feedback follows, this has no effect on our category learning. The semisupervised learning was also studied giving feedback after 25% of the trials. In this 25% semisupervised condition, participants failed to learn the task. This failure was not due to the low relative percentage of feedback trials in a block, because when the absolute number of trials was doubled (i.e., 25% feedback was maintained, but twice as many feedback trials were received), almost all participants were able to master the category structure. Thus, when given enough trials, even 25% feedback sufficed to learn the category structure. Again, this result suggests that the no-feedback trials have little impact on the initial learning process, but that learning is rather determined by the absolute number of feedback trials one receives.

An important question is whether these findings imply that people encounter objects, classify them and then simply delete this experience because no confirmation or correction is provided. If so, this would be in contrast to findings from machine learning where machines do use no-feedback trials to extend the knowledge gained from feedback trials. Remarkably, when supervised and semisupervised machine learning are compared, semisupervised machine learning can even achieve faster optimal performance (Chapelle, Scholkopf & Zien, 2006; Zhu & Goldberg, 2009). In machine learning, semisupervised learning is therefore the method used most often, also due to practical implications: semisupervised learning requires fewer feedback items, which must be annotated one by one by humans and therefore reduces time investment (Zhu & Goldberg, 2009).

Since Vandist et al. (2009) several human semisupervised category learning studies were conducted and the findings are not always consistent. In the study of Mcdonnell, Jew, and Gureckis (2012) no impact of the no-feedback trials was found. In this study the category label was shown only on some trials, but not on others. In the labelled trials all stimuli originated from one subset of the full category. In the unlabelled trials, the presented stimuli covered the full category. After the training phase, the category presentation of the participant was tested. Mcdonnell et al. (2012) found that a large weight was given to the labelled stimuli, making the unlabelled trials irrelevant.

In other studies, the impact of the no-feedback trials depended on the circumstances. First, semisupervised learning was observed in a speeded classification task but not if the responses were self-paced (Rogers, Kalish, Gibson, Harrison, & Zhu, 2010). Second, participants did use the no-feedback trials when the underlying categories were distinct and the gap between the categories was big. However, if the underlying categories were more ambiguous and the space between the categories was small but still existing, no effect of the no-feedback trials was found (Vong, Perfors, & Navarro, 2014). Third, Kalish, Zhu, and Rogers (2015) showed that the effect of the no-feedback trials depends on the age of the participants: young children (between 4 and 6 years old) were influenced by the no-feedback trials whereas no effects were found for older children (between 7 and 8 years old).

Finally, some studies did show that the no-feedback trials aided learning (Gibson, Rogers, Kalish, & Zhu, 2015; Kalish, Zhu, & Rogers, 2011; Lake & McClelland, 2011; Zhu, Gibson, Jun, Rogers, Harrison, & Kalish, 2010). However, all of these studies used unidimensional stimuli and a simple underlying category structure. Feedback was always given after a specific subset of stimuli. Based on these feedback trials only, a certain decision bound that splits the two categories can be expected. The stimuli of the no-feedback trials had a different mean and distribution than the stimuli of the feedback trials because the latter were extremes of the category. If participants take these no-feedback trials into account, the decision bound will be shifted. These studies showed that participants indeed use a shifted decision bound, implying that the no-feedback trials do have an impact on the learning process (Gibson et al., 2015; Kalish et al., 2011; Lake & McClelland, 2011; Zhu et al., 2010). Still, it is unlikely that in our daily life feedback is always provided after the same

subset of examples and other examples of the category are never followed by feedback. Contrarily, we believe that every example of a category can be followed by feedback.

Automaticity

Given the inconsistent research results on human semisupervised learning and the advantages of semisupervised learning in machines, we aim to further investigate the role of no-feedback trials in the human semisupervised category of learning. In the current study, we specifically investigate the role of no-feedback trials in developing *automaticity*. Once a learner reaches automaticity, cognitive or motor skills are executed faster, more accurately and require less attention in comparison to initial learners (Ashby & Crossley, 2012; Ashby, Turner, & Horvitz, 2010). Although various definitions and criteria of automaticity exist, researchers agree that automaticity is the result of extensive overtraining after the skilled behavior is well learned (Ashby et al., 2010; Hélie, Waldschmidt, & Ashby, 2010; Moors & De Houwer, 2006; Schneider & Chein, 2003; Nosofsky & Palmeri, 1997; Shiffrin & Schneider, 1977). Especially in categorization this is the main consensus since several studies showed that the criteria for automaticity proposed by Schneider and Shiffrin (1977) as no interference of dual task performance (Waldron & Ashby, 2001; Zeithamova & Maddox, 2006, 2007) and decrease in performance after switching keys (Ashby, Ell, & Waldron, 2003; Maddox, Bohil, & Ing, 2004b; Maddox, Glass, O'Brien, Filoteo, & Ashby, 2010a; Spiering & Ashby, 2008a) already apply for initial information-integration category learning. Consequently, in this article automaticity will be defined as the result of overtraining after good performance was obtained.

In cognitive science, two influential models of expertise presented in the literature are Logan's (1988) instance theory of automaticity and Rickard's (1997) component power laws theory. Both models assume that feedback remains essential through the development of automaticity and hence only make predictions about supervised learning, not about semi-supervised learning. In category learning, two important models explicitly deal with automaticity: the Exemplar-based random walk model (EBRW-model) of Nosofsky and Palmeri (1997) and the Subcortical Pathways Enable Expertise Development (SPEED model) of Ashby, Ennis, and Spiering (2007). The EBRW-model assumes that expertise develops as the number of stored exemplars increases. The more stored exemplars, the faster the response will be elicited. Since only exemplars followed by feedback will be stored (and activated as belonging to the category), supervised learning is essential and no clear predictions can be made about semisupervised learning based on this model. For semisupervised learning, the most relevant model about the development of automaticity in the information-integration structure is the SPEED-model. The SPEED-model assumes that categorization is regulated by two different pathways, a slow and a fast one. The *slow pathway* is supposed to originate in the visual cortex, passes by the basal ganglia and the thalamus, and ends in the premotor cortex. This is an indirect and subcortical pathway that includes at least four synapses. When positive feedback is given (after a correct categorization), dopamine in the striatum will be released and the active synapses will be strengthened. When negative feedback is given (after an incorrect answer) or no feedback at all, the strength of the synapses will be weakened. On the contrary, Ashby et al. (2007) state that the *fast pathway* only involves one synapse. This is a direct route from the visual association areas to the premotor cortex. In this cortical-cortical pathway, synapses are strengthened when there is both pre- and postsynaptic activation (i.e., Hebbian learning). This occurs independent of feedback.

In the SPEED-model, the development of categorization automaticity is defined as a gradual process. Early in learning, the main pathway is the slow subcortical pathway. As learning progresses, the fast cortical-cortical pathway becomes more salient and the subcortical pathway becomes less important. Eventually, experts only rely on the cortical-cortical pathway for their categorization (Ashby et al., 2007). Because this pathway is independent of feedback and the strength of the connections increases with the number of categorization responses, SPEED predicts that late in learning every type of trial will strengthen the connections, regardless of whether the categorization response is followed by feedback. Hence, late in learning, adding extra no-feedback trials to the training would have an impact on the development of automaticity and faster response times can be expected.

To test this hypothesis in Experiment 1, participants were trained in a supervised way on the information-integration structure for 2 days. After reaching an expert level with regards to the trained category structure, half of the participants continued to practice supervisedly on days 3 and 4. The other half practiced according to a 25% semisupervised scheme. Both groups of participants received an equal amount of feedback trials, implying that the semisupervised group received four times as many trials as the supervised group. Based on the premises of the SPEED-model, we hypothesize that the categorization in the semisupervised condition will be more automatic, as indexed by faster response times. If the no-feedback trials in semisupervised learning have no impact on the automaticity process, a similar level of automaticity (and thus equal response times) should be observed in both conditions.

Experiment 1

Method

Participants In total 34 participants (22 women, average age 21.4 years, *SD*=1.97, range=18–26 years) took part in the experiment in return for payment. If participants participated

for 2 days, they received 20 euro; if they participated for 5 days, the payment was 35–40 euro.

Design The experiment was organized on five consecutive days. In this way learning could benefit from between-session consolidation due to sleep (Censor, Karni, & Sagi, 2006; Stickgold, James, & Hobson, 2000a; Stickgold, Whidbee, Schirmer, Patel, & Hobson, 2000b; Stickgold & Walker, 2005). Participants were randomly divided into two conditions: the semisupervised condition (n = 19) and the supervised condition (n = 15). The first 2 days were equal for both conditions: on each of these days, 400 training trials were presented, divided into five blocks of 80 trials. Each trial was followed by feedback. The goal of this training phase was to master the category structure. Participants who achieved an average accuracy rate of 90% or more on the last two blocks of the second day were invited to the following phase. Participants who did not reach this expert level were excluded from the remainder of the experiment. On the third and fourth day, participants in the semisupervised condition were presented with 640 trials (eight blocks of 80 trials) and 25% of these trials were randomly followed by feedback. Participants in the supervised condition were shown 160 trials (two blocks of 80 trials) that where all followed by feedback. Consequently, the number of feedback trials was equal in both conditions. On the fifth and final day the test phase took place where all participants received 134 trials in one block. None of these trials were followed by feedback. It was decided to organize this "test" on a new day, ensuring that participants in both conditions were equally fit. Table 1 summarizes the differences between the two conditions.

Stimuli and apparatus The experiment was conducted using Tscope (Stevens, Lammertyn, Verbruggen, & Vandierendonck, 2006). Participants viewed the stimuli on a 17-in. LCD monitor with an 800 × 600 resolution at a distance of approximately one arm's length. The stimuli were gray 300 × 300 square-pixel Gabor patches, presented on a black screen. Two examples of Gabor patches can be seen in Fig. 2. In this study the "gratings" varied continuously on two dimensions: the spatial orientation and the spatial frequency. These dimensions are perceptually separable. The arbitrary stimulus coordinates were converted to physical units using the following transformations: spatial orientation was referred to in x degrees, with x varying between 0 to 100 degrees. Spatial frequency, expressed in cycles/pixel, was converted using $f(y)=0.01+(y/1500)$, with y varying between 0 and 100. These coordinates originated from the information-integration category structure, an example of which is displayed in Fig. 1. The optimal decision bound, which classifies the stimuli perfectly in two categories, is diagonal. In the semisupervised condition, participants viewed 2,080 stimuli in the first 4 days. These stimuli were generated by randomly sampling from two bivariate normal distributions, leading to 1,040 "A" stimuli and 1,040 "B" stimuli. Category A had a different mean to category B, but the variance and the covariance of both categories were the same. Due to random sampling, the optimal decision bound varied slightly from block to block, although the mean optimal decision bound in one day was $y=x$. The exact parameter values are shown in Table 2. As in the Ashby et al. (1999) study, the mean, the variance, and the covariance values were chosen in such a way that a linear decision bound based on one dimension would account for an accuracy of maximum 80%. The stimuli in the supervised condition were constructed in the same way as in the semisupervised condition, the only difference being that in this condition participants viewed 1,120 stimuli in the first 4 days, 560 "A" stimuli and 560 "B" stimuli. On the last day, all participants viewed 134 fixed stimuli, half of which were depicted from the stimuli A range; the other half originated from the stimuli B range, as can be seen on Fig. 3. Again, the optimal decision bound was $y=x$ and the category mean was identical to that of the previous days.

Table 1 Number of trials in each condition of Experiment 1

Condition	Type of trial	Day 1	Day 2	Day 3	Day 4	Test
Semisupervised	Feedback	400	400	160	160	0
	No-feedback	0	0	480	480	134
Supervised	Feedback	400	400	160	160	0
	No-feedback	0	0	0	0	134

Procedure All participants were tested individually in a dimly lit room. On the first 2 days, participants were informed that they would see stimuli that would appear one by one and that originated from two categories A and B. They were asked to respond by pressing A on the keyboard if they believed that the stimulus was an A and to press B when they believed the stimulus was a B. Participants were informed that they would receive feedback (i.e., the true category label) after each category response. They were also informed that it was possible to do the task without errors. Participants were told that at the end of day 2 the accuracy would be calculated and only the participants who achieved an average accuracy of 90% or more would be allowed to continue to the next days. At the end of each block, the percentage of the correct responses was printed on the screen. This percentage additionally encouraged them to do better in the next block. A trial started when a stimulus was projected in the middle of the screen until the participant responded. Immediately after the response the stimulus disappeared. The response time was self-paced. After the response, the feedback, consisting of correct/incorrect and the right category label, became visible at the bottom of the screen for

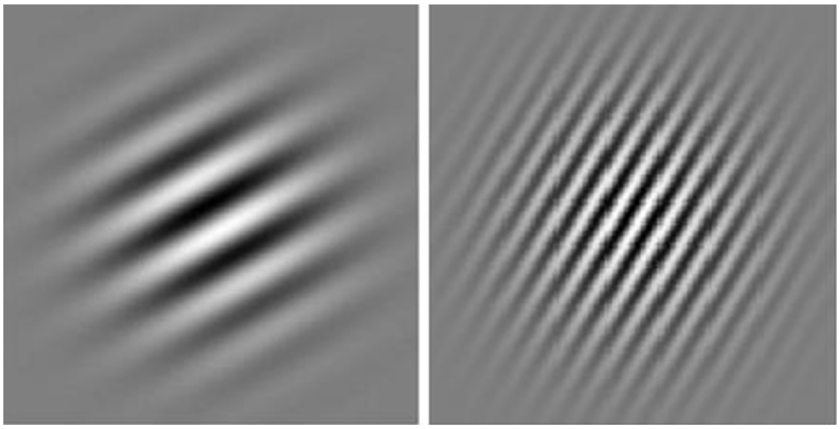

Fig. 2 Two examples of Gabor patches

1,500 ms. After that a new trial started. The procedure on days 3 and 4 was similar, except that in the semisupervised condition participants were informed that some trials would be followed by feedback and other trials not. In the no-feedback trials the stimulus disappeared immediately after the response and the screen remained blank for 1,500 ms. Afterwards a new trial started. Hence, there is no difference in post-response events on feedback trials and no-feedback trials except on the appearance of the feedback on the screen. Again, in both conditions participants were informed that it was possible to obtain maximum accuracy and they were encouraged to achieve this. On day 5 participants were informed that they would see similar stimuli to those in the preceding days and that feedback would no longer be given. In contrast to the first 4 days, participants were urged to respond as quickly as possible. After a response was given, the stimulus disappeared and the screen remained blank for 1,500 ms, after which a new trial started. To encourage the participants to respond as fast as possible, the stimulus disappeared when a response was not given within a time limit of 1,800 ms. In this case, the message "Respond faster" was shown during the subsequent intertrial interval of 1,500 ms. The trials in which the participant responded too slowly were presented again at the end of the block. Thus, for each participant we collected 134 valid categorization responses on day 5.

Table 2 Parameter values that define the categories of Experiment 1

Category	A	B
Mean (frequency)	40	60
Mean (orientation)	60	40
SD (frequency and orientation)	11.88	11.88
Correlation	0.99	0.99
Average optimal slope	1	
Average optimal intercept	0	

Results

Selection of participants

Before analyzing the response time patterns in both conditions, it was essential to ensure that the participants mastered the category structure at the end of day 2. Therefore accuracy and model-based analyses were performed. High accuracy rates indicate that the participant made few errors. Nevertheless, it is still unclear whether these errors were just random mistakes or systematic faults. The model-based analyses are a necessary complement to the accuracy. These models were calculated based upon the responses of the participant on the last two blocks of day 2. For each model the corresponding BIC score were calculated. The best fitting

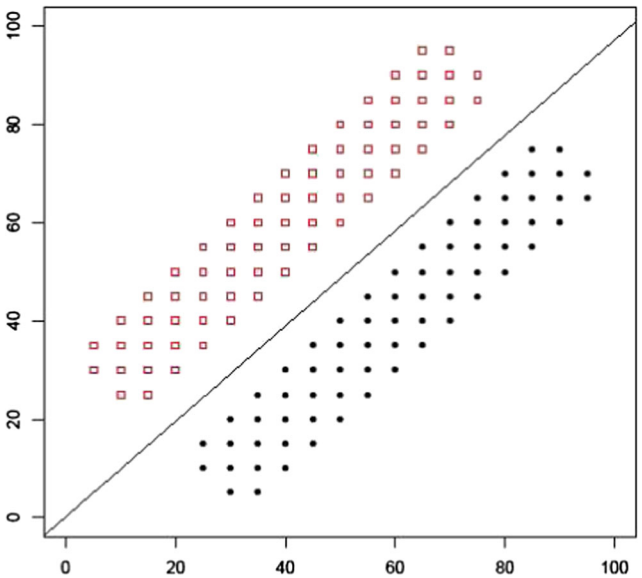

Fig. 3 An example of the information-integration category structure used in the test (= day 5) in Experiment 1. The "A" stimuli are shown in squares, the "B" stimuli in solid circles. The decision bound that divides the two categories perfectly is shown in black. The Y-axis is the Orientation dimension, the X-axis the Frequency dimension

model was the model with the lowest BIC score. This model is supposed to correspond to the strategy that the participant used to solve the categorization task. The strategy that matched perfect performance in this experiment is called the optimal decision bound. Combining both criteria rules out the possibility that a participant increased the accuracy during the experiment, but this improvement was not reflected in the model-based analyses. This can be the case if the errors are systematic and a strategy other than the optimal decision bound was preferred. Therefore, only participants that passed both criteria were retained for further analyses.

Criterion 1: High accuracy

Participants could achieve perfect accuracy when using the optimal decision bound (i.e., by integrating the information from the two stimulus dimensions at some predecisional stage). On the other hand, using a unidimensional decision rule, the accuracy could never exceed 80%. This implies that participants with an achieved accuracy of more than 80% probably adopted a (suboptimal) information-integration decision rule. However, since our study aimed at studying automaticity after becoming an expert learner, category learning was considered successful when an average performance of at least 90% was obtained. As can be seen in Table 3, 11 participants did not pass this criterion and were therefore excluded from further analyses.

Criterion 2: Optimal decision bound Figures 1 and 2 in the Supplementary Materials show the actual responses during the last two blocks of day 2 for each of the participants retained after applying Criterion 1. These responses (i.e., whether a stimulus belongs to category A or category B) form the basis on which the individual decision bounds were calculated. Four different types of models were fit to each participant's response (see the Appendix for details). These models were

Table 3 Mean accuracy (%) and model-based analyses (BIC scores) of the last two blocks of day 2 (i.e., blocks 9 and 10) for every participant of Experiment 1

Condition	Participant	GLC	DIM-O	DIM-F	GCC	Mean accuracy (%)	Continued to day 3
Semisupervised	1	120.76	210.91	151.31	113.88	86.88	No
	2	36.84	176.66	118.31	61.79	98.13	Yes
	3	18.44	148.80	167.54	52.90	99.38	Yes
	4	103.40	184.60	179.89	75.66	91.88	No
	5	163.81	173.48	225.16	164.46	78.75	No
	6	81.82	148.98	190.57	105.74	93.75	Yes
	7	115.00	215.45	143.54	126.24	86.88	No
	8	15.23	146.18	150.06	67.04	100.00	Yes
	10	106.76	240.21	104.52	109.57	78.13	No
	11	67.49	202.81	121.77	78.83	94.38	Yes
	12	75.97	160.99	170.85	78.36	95.00	Yes
	13	45.92	144.50	171.39	85.92	97.50	Yes
	15	113.02	233.09	113.54	118.61	81.88	No
	16	45.21	182.43	146.23	95.33	96.88	Yes
	17	72.62	166.22	173.66	110.02	95.63	Yes
	18	187.60	230.57	191.50	187.31	73.75	No
	19	47.01	117.47	202.99	66.36	95.38	Yes
	20	33.64	143.68	163.89	74.42	98.75	Yes
	26	154.79	241.89	152.85	154.58	74.38	No
Supervised	1	96.75	176.61	173.71	97.82	93.13	Yes
	2	51.72	164.90	156.70	79.18	97.50	Yes
	3	36.10	156.05	165.01	89.58	98.13	Yes
	4	58.12	165.86	161.56	83.26	96.88	Yes
	5	115.82	160.44	199.81	117.14	89.38	No
	6	108.42	215.85	128.97	127.23	88.13	No
	7	90.37	185.83	160.68	106.73	92.50	Yes
	8	116.05	204.53	159.70	124.89	89.38	No
	9	49.23	171.32	134.58	86.52	96.25	Yes
	10	137.92	227.78	148.88	142.17	83.13	No
	11	63.21	159.75	180.47	103.53	96.25	Yes
	12	49.95	150.89	159.37	62.27	96.88	Yes
	14	80.05	168.53	176.18	90.21	95.00	Yes
	15	89.64	144.01	206.52	102.48	91.88	No
	18	98.25	149.18	194.89	107.51	91.88	Yes

The best fitting model is underlined. Only participants who reached an accuracy of minimal 90% and had the best fitting decision bound model based on the GLC, continued on to day 3 of the experiment

DIM-O unidimensional classifier based on the orientation, *DIM-V* unidimensional classifier based on the frequency, *GCC* General Conjunctive Classifier, *GLC* General Linear Classifier

introduced by Ashby and Gott (1988) and Ashby and Maddox (1993). Three models, namely the horizontal unidimensional model (DIM-O), the vertical unidimensional (DIM-V) and the general conjunctive classifier (GCC), are rule-based. If participants adopted one of these category decision strategies, category learning failed. The last model, the general linear classifier (GLC), is an information-integration model. Only with this category decision strategy perfect accuracy could be obtained. Consequently, if the general linear classifier was used and this decision bound fell in between the two categories, learning was successful. The model parameters were estimated using the method of maximum likelihood. To select the best-fitting model to the data, the model with the smallest Bayesian Information Criterion (BIC) was selected. The BIC penalizes according to the number of free parameters. BIC is defined as BIC = $r\ln N - 2\ln L$, where r is the number of free parameters, N is the sample size and L is the likelihood of the model given the data (Schwarz, 1978). The BIC values of the four models for each participant are presented in Table 3. In the semisupervised condition, all participants except participant 4 favored the general linear classifier. Hence, participant 4 was excluded from further analyses. As can be seen in Fig. 1, all optimal decision bounds fell between the two categories. In the supervised condition, all participants favored a strategy based on the general linear classifier. As can be seen in Fig. 2, all optimal decision bounds fell between the two categories except for participant 15. Hence, participant 15 was excluded from further analyses. As a result, the final sample used in the subsequent analyses consisted of 21 participants (n=11 semisupervised and n=10 supervised), the average age was 21.3 years (*SD*=1.88, range 18–24 years), and 15 of them were women.

In the following sections, four types of analyses are described: accuracy and model-based analyses to define the strategy used by the participant, response time (RT) analyses, and the speed-accuracy trade-off analyses. The response time analyses were studied to compare the semisupervised learning process to the supervised learning process.

Accuracy analysis Figure 4 shows the average percentage of correct responses and the 95% confidence intervals on each block of trials received during the first 4 days for the supervised and semisupervised condition separately. In the semisupervised condition, the accuracy was based on the feedback trials only. Eighty feedback trials were grouped into a block to facilitate comparison to the supervised condition. As expected, the learning process was similar in both conditions. The mean accuracy increased from an average of 73% (*SD*=11.06) in the first block for the semisupervised condition and an average of 75% (*SD*=11.72) for the supervised condition to almost perfect accuracy in the last block of day 2 (97%, *SD*=2.76 and 95%, *SD*=3.62, respectively). During the blocks on days 3 and 4, the mean accuracy remained high in both

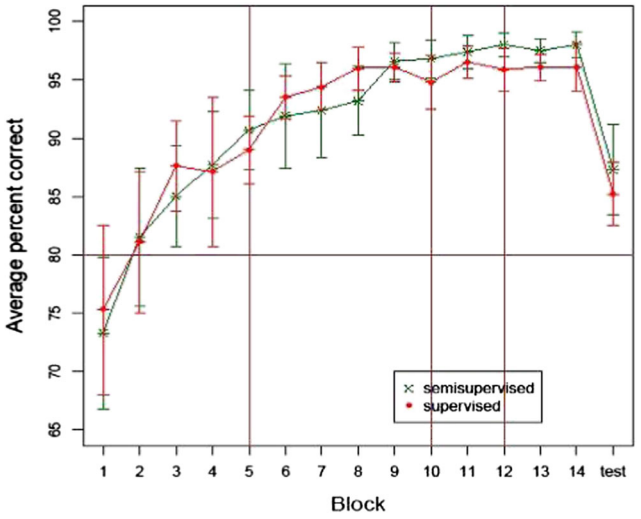

Fig. 4 Mean accuracy (%) along with the 95% confidence intervals by block for all participants in the Semisupervised and Supervised Condition of Experiment 1 from day 1 (blocks 1–5), day 2 (blocks 6–10), day 3 (blocks 11–12), day 4 (blocks 13–14), and day 5 (test). Only the participants who reached a performance of minimum 90% at the end of day 2 and revealed a decision bound based on the optimal decision bound were included in the analyses. In the Semisupervised condition, the accuracy for every 80 feedback trials is used

conditions. In the semisupervised condition, mean accuracy on the last response block on day 3 was 98% (*SD*=1.80) and 98% on day 4 (*SD*=1.89). Similarly, in the supervised condition mean accuracy on the last block was 96% (*SD*=2.95) on day 3 and 96% (*SD*=3.41) on day 4. A repeated measures ANOVA was conducted to determine whether the mean accuracy on the last two blocks differed depending on the day (4 levels: day 1, 2, 3, and 4) and condition (2 levels: supervised and semisupervised). Not surprisingly, there was a main effect of day, $F(3,17)=13.97$, $p<.001$, $\eta_p^2=.71$, indicating that the accuracy increased during the succeeding days. Paired sample *t*-tests using the Bonferroni correction for multiple comparisons showed that in comparison to day 1, mean accuracy was significantly higher on days 2, 3 and 4 (resp. $t(20)=6.41$, $p<.001$; $t(20)=6.80$, $p<.001$; $t(20)=7.02$, $p<.001$). There was no main effect of condition, $F(1,19)=1.71$, $p=.21$, $\eta_p^2=.08$, nor an interaction between day and condition ($F<1$, $p=.99$, $\eta_p^2=.007$), suggesting that the accuracy in both conditions increased similarly across days. The accuracy on the test (day 5), where the speed of responding was stressed, was lower in both conditions compared to the accuracy reached at the end of day 4: 87% (*SD*=6.53) in the semisupervised condition and 85% (*SD*=4.37) in the supervised condition. This difference was significant for the semisupervised condition, $t(10)=5.42$, $p<.001$, $d=1.64$, and for the supervised condition $t(9)=7.46$, $p<.001$, $d=2.36$. Finally, an independent sample *t*-test revealed that there was no difference in accuracy between the two conditions on the fifth day, $t(19)=0.91$, $p=.37$, $d=0.40$.

Model-based analysis Table 4 shows the four model fits on day 4. In both conditions, all participants preferred a decision bound based on the general linear classifier, indicating successful learning. Figures 3 and 4 in the Supplementary Materials show the actual responses during the last two blocks of day 4 for each participant. Table 5 presents the model fits on the test day (i.e., day 5): in the semisupervised condition, most participants chose a strategy based on the optimal decision bound, except for participants 6, 12, and 19. For these participants a strategy based on the general conjunctive classifier fitted slightly better than the optimal decision bound. Figures 5 and 6 in the Supplementary Materials present the responses and the best fitting decision bound for every participant during the test day. In the supervised condition, most participants preferred the optimal decision bound, except for participants 1 and 18. For participant 1, the BIC score of the general conjunctive classifier was slightly lower than the general linear classifier. Participant 18 clearly preferred a strategy based on the general conjunctive classifier.

Analysis of the response times The mean response times (RTs) along with the 95% confidence intervals for the semisupervised and supervised condition from day 1 to day 4 are presented in Fig. 5. These RTs were calculated on the last two blocks of each day. For day 1, the mean RTs in the supervised condition was 944 ms (*SD* =145.48) whereas the mean RTs in the semisupervised condition was 858 ms (*SD*=131.22). Importantly for this investigation is that participants were equally fast in the last two blocks of day 2 (semisupervised mean RT=785 ms, *SD*=90.10 and supervised mean RT=801 ms, *SD*=122.22). This was confirmed by an independent sample *t*-test, t (19)=0.35, p=.73, d=*0.15*. On days 3 and 4, the mean RTs slowly decreased in the semisupervised condition to, respectively, 794 ms (*SD*=106.53) and 719 ms (*SD*=128.59). This decrease in mean RTs was also observed in the supervised condition: 719 ms (*SD*=108.98) on day 3 and 724 ms (*SD*=126.85) on day 4. A repeated measures ANOVA was conducted to determine whether the mean RTs on the last two blocks differed depending on the day (4 levels: day 1, 2, 3, and 4) and condition (2 levels: supervised and semisupervised). Not surprisingly, there was a main effect of day, $F(3,17)=7.12$, p=.003, η_p^2=*.56*, indicating that the RTs

Table 4 Mean accuracy (%) and model-based analyses (BIC scores) on the last two blocks of day 4 (i.e., blocks 7–8 semisupervised condition; blocks 1–2 supervised condition) for every participant of Experiment 1

Condition	Participant	GLC	DIM-O	DIM-F	GCC	Mean accuracy (%)
Semisupervised	2	37.21	152.60	156.77	83.20	98.75
	3	40.21	142.25	169.50	68.89	100.00
	6	67.27	157.05	184.21	80.58	96.88
	8	15.23	151.83	140.60	64.52	97.50
	11	15.23	150.92	162.51	33.45	98.75
	12	41.56	161.91	156.19	82.15	95.00
	13	28.17	153.32	155.14	58.71	98.75
	16	75.25	162.52	157.88	81.30	95.63
	17	25.30	159.73	127.14	62.00	98.13
	19	57.56	138.92	183.19	68.14	96.25
	20	15.23	140.19	145.79	61.59	99.40
Supervised	1	44.37	270.08	164.14	82.63	97.50
	2	47.95	258.59	160.92	80.58	97.50
	3	80.20	256.55	166.29	99.31	94.38
	4	64.59	254.25	137.36	96.71	95.00
	7	91.93	262.33	160.30	108.01	93.13
	9	58.14	258.02	158.47	73.30	96.88
	11	35.34	259.01	143.44	80.13	98.75
	12	40.75	265.04	166.17	85.23	98.13
	14	64.73	258.52	164.25	89.22	96.25
	18	81.47	262.97	189.35	102.19	93.75

The best fitting model is underlined

DIM-O unidimensional classifier based on the orientation, *DIM-V* unidimensional classifier based on the frequency, *GCC* General Conjunctive Classifier, *GLC* General Linear Classifier

Table 5 Mean accuracy (%) and model-based analyses (BIC scores) on all trials of day 5 for every participant of Experiment 1

Condition	Participant	GLC	DIM-O	DIM-F	GCC	Mean accuracy (%)
Semisupervised	2	66.35	166.04	173.69	113.15	94.78
	3	57.34	161.63	175.74	99.05	95.52
	6	146.23	178.73	186.08	141.14	80.60
	8	111.43	154.48	193.88	122.96	84.33
	11	69.64	179.10	160.71	93.81	92.54
	12	118.17	171.33	183.52	111.44	85.07
	13	148.39	159.83	191.96	148.86	73.88
	16	98.29	162.05	185.89	121.93	88.81
	17	91.97	179.02	168.41	116.15	89.55
	19	116.41	169.27	187.80	111.35	84.33
	20	58.78	149.03	184.81	92.53	91.79
Supervised	1	130.18	179.16	182.94	127.17	84.33
	2	103.71	202.83	154.30	120.48	85.83
	3	125.65	168.79	189.29	139.90	83.58
	4	127.66	140.63	197.70	139.62	76.87
	7	69.25	174.38	166.44	109.05	92.54
	9	100.43	171.17	171.63	103.40	85.82
	11	64.10	145.70	181.16	91.09	89.55
	12	95.96	157.49	188.50	123.50	87.31
	14	104.43	153.38	174.07	130.30	80.60
	18	102.03	149.89	197.79	74.00	85.07

The best fitting model is underlined

DIM-O unidimensional classifier based on the orientation, *DIM-V* unidimensional classifier based on the frequency, *GCC* General Conjunctive Classifier, *GLC* General Linear Classifier

decreased during the succeeding days. There was no main effect of condition, $F<1$, $p=.84$, $\eta_p^2=.002$, but the interaction between day and condition reached significance $F(3,17)=3.92$, $p=.027$, $\eta_p^2=.41$. Post hoc paired-sample *t*-tests, adjusted by the Bonferroni correction for multiple comparisons, indicated that in the semisupervised condition participants did not speed up by day, none of the paired sample *t*-tests was significant, all $p>.09$. In the supervised condition, responses on later days were all faster in comparison to day 1 (all $p < .02$). None of the other comparisons were significant (all p>.06). The decrease in RTs is thus only present in the supervised condition.

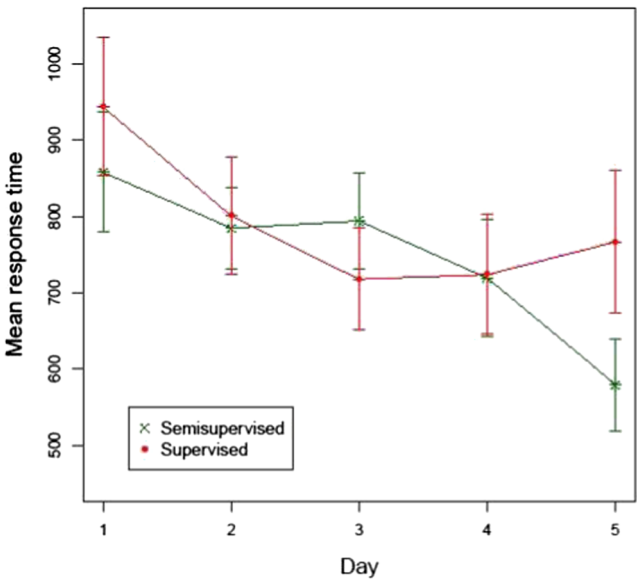

Fig. 5 Mean response times (ms) and 95% confidence intervals for the semisupervised and the supervised condition of Experiment 1 calculated on the last two blocks of each day

Decisive to test our hypotheses was the difference in RTs on day 5. In the semisupervised condition the mean RT was 579 ms ($SD=104.08$) whereas the mean RT in the supervised condition was 767 ms ($SD=153.20$). An independent sample *t*-test confirmed that this difference in RTs on the test day was significant, $t(19)=3.32$, $p=.004$, $d=1.45$. Participants in the semisupervised condition responded significantly faster than participants in the supervised condition. Paired sample *t*-tests showed that in the semisupervised condition, participants responded significantly faster on day 5 compared to day 4, $t(10)=4.71$, $p=.001$, $d=1.42$ whereas participants in the supervised condition responded equally fast on days 4 and 5, $t(9)=1.39$, $p=.20$, $d=0.44$.

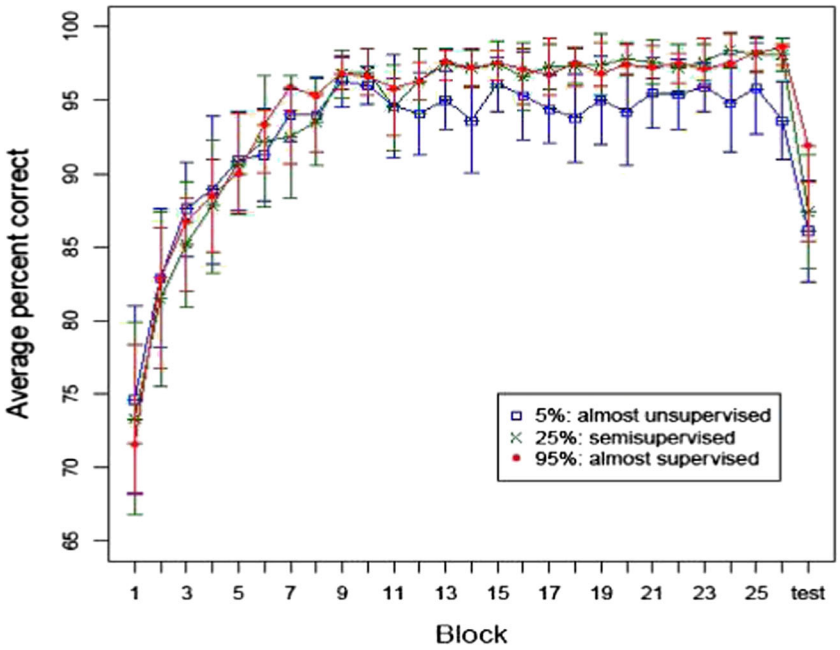

Fig. 6 Mean accuracy (%) and 95% confidence intervals by block for all participants in the Almost supervised and Almost Unsupervised Condition of Experiment 2 and Semisupervised Condition of Experiment 1 from day 1 (blocks 1–5), day 2 (blocks 6–10), day 3 (blocks 11–18), day 4 (blocks 19–26) and day 5 ("t"). Only the participants who reached a performance of minimum 90% at the end of day 2 and revealed a decision bound based on the optimal decision bound were included in the analyses. All trials (feedback and no-feedback trials) were adopted in the analyses

Speed-accuracy trade-off On the fifth day, participants were instructed to respond as fast as possible. This might result in a speed-accuracy trade-off (SAT): participants gave up decision accuracy in favor of decision speed (see Heitz, 2014).The speed-accuracy trade-off was calculated for each condition. Since data points are limited and errors are rare, the SAT is calculated by the Pearson correlation between the mean RT and the mean accuracy rate (Heitz, 2014). In the supervised condition, there was no SAT-effect, r=-.21, p=.57. In the semisupervised condition, there was a SAT-effect, r=.61, p=.046, implying that the faster participants responded, the more errors they made.

Discussion

The objective of Experiment 1 was to test the hypothesis based on the SPEED-model that late in learning, when automaticity develops, participants benefit from semisupervised learning, resulting in faster RTs. Therefore, participants were trained during two days until a certain expertise was gained and then either received feedback on all trials (supervised condition), or on 25% of the trials (semisupervised condition) for the next 2 days. On days 3–4 both conditions received an equal amount of feedback trials but the total number of trials differs: in the semisupervised condition, participants categorized a quadruple of trials compared to the supervised condition. The results clearly showed that the mean RTs on the test (day 5) were significantly faster in the semisupervised condition than in the supervised condition. The participants in the semisupervised condition revealed more automatic behavior than the participants in the supervised condition. Importantly, this difference in mean RTs on day 5 was not observed on day 2, ruling out the possibility that participants in the semisupervised condition were always faster. On day 5 a SAT-effect occurred in the semisupervised condition: participants who tended to respond fast, also made more errors. This was not the case in the supervised condition. Note that the mean accuracy was similar on day 5 in both conditions: even though semisupervised participants sacrificed accuracy for response times, they still performed at the same level with regards to accuracy as supervised participants.

There are three possible explanations for these findings. The first is that semisupervised learning does have an impact late in learning when automaticity develops and that it leads to faster RTs. Second, confirming the SPEED-model, it is possible that the higher number of trials in the semisupervised condition is accountable for the faster RTs. On days 3 and 4 participants in the semisupervised condition responded to a fourfold number of trials. The SPEED-model postulates that late in learning the fast pathway is dominant. This pathway is assumed to be independent of feedback. According to the SPEED-model, the multiple repetitions in the semisupervised condition lead to faster response times, regardless whether or not these trials are followed by feedback. Third, the results may have been influenced by a confound. The participants in the supervised condition never experienced no-feedback

trials before day 5 and this inexperience might have slowed their performance. In Experiment 2 these explanations were addressed.

Experiment 2

Experiment 2 again examines whether late in learning the nature of feedback (i.e., feedback on every trial, occasional feedback, or no-feedback) has an impact on the development of automaticity, indicated by faster RTs. In order to do this without the confounds present in Experiment 1, two control conditions are run and compared to the semisupervised condition of Experiment 1: an almost supervised condition, in which 95% of the trials are followed by feedback and an almost unsupervised condition, in which 5% of the trials are followed by feedback.[1] Instead of using fully supervised and unsupervised control conditions, "almost" supervised and unsupervised conditions are purposefully chosen so that all participants have experienced no-feedback trials prior to day 5, ruling out the possibility that the novelty of no-feedback trials on day 5 influences the RTs of the supervised condition. The total number of trials in these two new conditions corresponds to the semisupervised condition of Experiment 1, to test the hypothesis that just the higher number of trials in the semisupervised condition of Experiment 1 led to faster RTs on day 5. If the total number of trials is indicative for the development of automaticity, similar RTs are expected in these two new conditions to those in the semisupervised condition of Experiment 1. This is the outcome predicted by the SPEED-model (Ashby et al., 2007). Contrarily, when the RTs in the two new conditions differ from the semisupervised condition of Experiment 1, this effect will be due to the different amount of feedback trials. If RTs in the *almost supervised condition* are faster than the semisupervised condition, either the higher amount of feedback trials in the supervised condition aids automaticity, or the higher amount of no-feedback trials in the semisupervised condition slows down learning. On the other hand, if RTs in the almost supervised condition are slower than the semisupervised condition, semisupervised learning aids the development of automaticity, despite the lower amount of feedback trials.

For the *almost unsupervised condition*, the SPEED-model predicts similar RTs to those in the semisupervised condition of Experiment 1, since only expert participants are selected. When RTs do differ from the semisupervised condition, this will provide us with insight into the minimum amount of feedback trials needed to successfully develop automaticity.

[1] Technically, the almost supervised condition and the almost unsupervised condition are semisupervised conditions as well, since these conditions consist of supervised and unsupervised trials. However, for clarity, when we refer to the "semisupervised condition" we intend the 25% semisupervised condition throughout the manuscript.

Method

Participants, design, stimuli, apparatus and procedure In total 38 participants (28 women, average age 20.74 years, *SD*=3.18, range=18–30 years) took part in the experiment in return for payment. The background of the participants was similar to the participants of Experiment 1. Also, the time of testing in the academic year was comparable. If participants participated for 2 days, they received 10 euro; if they participated for 5 days, the payment was 30 euro. Participants were randomly divided into two conditions: the almost supervised condition (n=19) and the almost unsupervised condition (n=19). The organization of Experiment 2 was identical to the semisupervised condition of Experiment 1 except on the third and fourth days. In the almost supervised condition, participants were presented with 640 trials (eight blocks of 80 trials) and 95% of these trials were randomly followed by feedback, resulting into 608 feedback trials on days 3 and 4. In the almost unsupervised condition participants also received 640 trials (eight blocks of 80 trials) but only 5% of these trials were randomly followed by feedback, resulting into 32 feedback trials on days 3 and 4. Note that the total number of trials in the semisupervised condition of Experiment 1 is the same as the total number of trials in the two conditions of Experiment 2. Table 6 presents the conditions schematically.

Results

Selection of participants

As in Experiment 1, accuracy and model-based analyses were performed to ensure that the participants mastered the category structure at the end of day 2.

Criterion 1: High accuracy Recall that participants could achieve perfect accuracy in this task. As in Experiment 1, the criterion was an average performance of at least 90% on the last two blocks of day 2. As can be seen in Table 7, ten participants did not pass this criterion and were therefore excluded from further analyses.

Criterion 2: Optimal decision bound Figures 7 and 8 in the Supplementary Materials show the actual responses during the last two blocks of day 2 for each participant who passed the accuracy criterion. As in Experiment 1, these responses were used to calculate the individual decision bounds of four different models and the corresponding BIC scores. The model with the lowest BIC score is presumably the strategy that the participant adopted in the last two blocks of day 2. Only participants who favored the general linear classifier model with a decision bound falling between the

Table 6 Number of trials in each condition of Experiment 2 and the semisupervised condition of Experiment 1

Condition	Type of trial	Day 1	Day 2	Day 3	Day 4	Test
Semisupervised Experiment 1	Feedback	400	400	160	160	0
	No-feedback	0	0	480	480	134
	Total	400	400	640	640	134
Almost supervised	Feedback	400	400	608	608	0
	No-feedback	0	0	32	32	134
	Total	400	400	640	640	134
Almost unsupervised	Feedback	400	400	32	32	0
	No-feedback	0	0	608	608	134
	Total	400	400	640	640	134

two categories were retained. In both conditions all remaining participants favored a strategy based on the general linear classifier. As can be seen in Figs. 7 and 8 in the Supplementary Materials, all optimal decision bounds fell between the two categories. As a result, the final sample used in the subsequent analyses consisted of 28 participants (19 women; average age of 20.9 years, *SD*=3.24, range=18–30 years years; n= 15 in the almost supervised condition and n=13 in the almost unsupervised condition).

As in Experiment 1, four types of analyses are reported: accuracy, model-based, response time analyses, and SAT analyses. The response time analyses were used to compare the learning process of the almost supervised and almost unsupervised conditions of Experiment 2 to the semisupervised condition of Experiment 1.

Accuracy analysis Figure 6 shows the average percentage of correct responses and the 95% confidence intervals on each block of trials received during the first 4 days for each of the conditions (almost supervised, almost unsupervised, and the semisupervised condition of Experiment 1) separately. In all conditions the accuracy was based on all trials (feedback and no-feedback trials). During the first 2 days (blocks 1–8) the learning process was similar in the three conditions. The mean accuracy increased from an average of 72% (*SD*=9.93) in the first block for the almost supervised condition, an average of 75% (*SD*=11.76) for the almost unsupervised condition, and an average of 73% (*SD*=11.06) for the semisupervised condition to almost perfect accuracy in the last block of day 2 (97%, *SD*=3.11, 96%, *SD*=2.35 and 97%, *SD*=2.76, respectively). During the blocks on days 3 and 4, the mean accuracy was almost perfect in the almost supervised condition: the mean accuracy on the last block on day 3 was 98% (*SD*=2.59) and 99% (*SD*=1.48) on day 4. In the almost unsupervised condition the mean accuracy was also high: the mean accuracy on the last block on day 3 was 94% (*SD*=5.54) and 94% (*SD*=4.92) on day 4. For the semisupervised condition of Experiment 1, the mean accuracy on the last block of day 3 was 97% (*SD*=2.00) and 97% (*SD*=2.00) on day 4. A repeated measures ANOVA was conducted to determine whether the mean accuracy on the last two blocks differed depending on the day (4 levels: days 1, 2, 3, and 4) and condition (three levels: almost supervised, almost unsupervised, and semisupervised). Not surprisingly, there was a main effect of day, $F(3,34)=22.67$, $p<.001$, $\eta_p^2=.67$, indicating that the accuracy significantly increased during the succeeding days. Post hoc paired-sample *t*-tests, adjusted by the Bonferroni correction for multiple comparisons, indicated that in comparison to day 1, mean accuracy was significantly higher on days 2, 3, and 4 (resp. $t(38)=7.79$, $p<.001$; $t(38)=5.60$, $p<.001$; $t(38)=7.18$, $p<.001$). There was no main effect of condition, $F(2,36)=1.60$, $p=.22$, $\eta_p^2=.08$ nor an interaction between day and condition ($F(6,68)=1.60$, $p=.16$ $\eta_p^2=.12$, suggesting that the increase in accuracy across days was similar in the three conditions.

The accuracy on the test (day 5), where the speed of responding was stressed, was lower in all conditions compared to the accuracy reached at the end of day 4, the average difference was -6.7% (*SD*=3.35) in the almost supervised condition, -7.5% (*SD*=7.47) in the almost unsupervised condition, and -10.69% (*SD*=6.97) in the semisupervised condition. Paired sample *t*-tests showed that these differences were significant: accuracy significantly dropped between the last block of day 4 and day 5, $t(14)=7.69$, $p<.001$, $d=2.06$, for the almost supervised condition, $t(10)=5.08$, $p<.001$, $d=1.53$ for the semisupervised condition, and $t(12)=3.62$, $p=.003$, $d=1.00$ for the almost unsupervised condition. Finally and crucially, a one-way ANOVA was conducted to determine whether the mean accuracy on day 5 differed between the almost supervised, almost unsupervised, and the semisupervised condition of Experiment 1. This was the case: $F(2,36)=4.18$, $p=.02$, $\eta_p^2=.19$. Independent sample *t*-tests corrected by the Bonferroni correction for multiple comparisons showed that this effect is due to the significant difference in accuracy between the almost supervised condition (92%, *SD*=4.08) and the almost unsupervised condition (86%, *SD*=6.40), $t(26)=2.94$, $p=.02$. The accuracy in the semisupervised

Table 7 Mean accuracy (%) and model-based analyses (BIC scores) of the last two blocks of day 2 (i.e., blocks 9 and 10) for every participant of Experiment 2

Condition	Participant	GLC	DIM-O	DIM-F	GCC	Mean accuracy (%)	Continued to day 2
95%	6	43.92	278.61	158.77	83.84	98.13	Yes
	8	34.52	262.25	164.44	88.15	98.75	Yes
	10	120.39	259.54	132.87	127.56	83.13	No
	11	90.23	274.58	164.66	92.92	93.75	Yes
	12	101.52	258.13	187.90	125.69	91.88	Yes
	13	143.06	242.65	144.00	143.52	78.75	No
	14	68.35	264.54	187.13	85.75	95.63	Yes
	15	64.38	258.93	169.46	98.06	96.25	Yes
	16	104.53	236.14	134.35	84.65	86.88	No
	17	28.63	252.74	151.33	64.86	98.75	Yes
	18	15.23	272.62	125.92	79.57	100.00	Yes
	19	59.41	264.04	157.14	95.63	96.88	Yes
	20	33.80	261.05	153.41	157.96	98.75	Yes
	21	39.00	257.39	155.33	91.06	98.13	Yes
	22	55.66	274.62	147.26	89.00	96.88	Yes
	30	87.82	274.10	174.76	105.20	93.75	Yes
	31	61.31	273.55	167.50	66.92	96.25	Yes
	33	66.43	260.26	164.10	100.90	96.25	Yes
	34	112.86	260.82	135.03	132.94	85.63	No
5%	1	122.08	256.69	157.20	133.90	87.50	No
	2	52.42	258.24	171.21	96.19	97.50	Yes
	3	15.23	259.17	133.39	139.17	100.00	Yes
	4	167.87	170.41	168.98	172.67	54.38	No
	5	15.23	262.82	134.55	87.14	99.38	Yes
	9	47.62	255.85	134.61	98.70	96.88	Yes
	23	101.30	255.41	161.08	118.23	91.88	Yes
	24	47.76	280.27	139.44	93.26	96.88	Yes
	25	86.81	270.69	192.59	101.06	92.50	Yes
	26	120.97	267.43	176.05	130.84	88.75	No
	27	151.78	253.21	154.40	157.07	80.63	No
	28	77.48	279.11	174.92	97.19	95.00	Yes
	32	83.84	268.02	180.18	120.27	94.38	Yes
	35	136.31	277.26	203.32	128.85	86.88	No
	36	73.27	259.55	170.94	103.65	95.00	Yes
	37	50.27	278.40	181.38	64.42	95.63	Yes
	38	158.81	250.97	205.01	156.82	82.50	No
	39	36.39	253.73	135.00	69.26	97.50	Yes
	40	52.58	258.91	154.59	74.19	97.50	Yes

The best fitting model is underlined. Only participants who reached an accuracy of minimal 90% and had the best fitting decision bound model based on the GLC, continued on to day 3 of the experiment

DIM-O unidimensional classifier based on the orientation, *DIM-F* unidimensional classifier based on the frequency, *GCC* General Conjunctive Classifier, *GLC* General Linear Classifier

condition (87%, *SD*=6.53) did not differ significantly from the almost unsupervised condition ($t(22)=.50$, $p=1$) and, most crucially, did not differ significantly from the almost supervised condition ($t(24)=2.19$, $p=.11$).

Model-based analysis Table 8 shows the four model fits on day 4. In the almost supervised condition, all participants preferred a decision bound based on the general linear classifier, indicating successful learning. In the almost unsupervised

Table 8 Mean accuracy (%) and model-based analyses (BIC scores) on the last two blocks of day 4 for every participant of Experiment 2

Condition	Participant	GLC	DIM-O	DIM-F	GCC	Mean accuracy (%)
Almost supervised	6	15.23	140.25	158.76	53.95	100.00
	8	53.02	162.44	145.39	88.09	97.50
	11	36.08	159.57	164.48	86.37	98.75
	12	24.06	140.65	160.16	66.73	99.38
	14	21.01	158.51	162.55	71.50	99.38
	15	82.40	198.44	132.80	107.23	93.13
	17	44.84	156.04	169.13	75.68	97.50
	18	23.55	133.00	141.50	57.27	99.38
	19	15.23	128.30	169.71	70.16	99.38
	20	15.23	148.01	134.59	63.61	100.00
	21	27.92	144.63	158.16	66.75	98.75
	22	32.74	122.05	177.36	60.28	98.75
	30	25.60	146.63	171.97	70.64	99.38
	31	43.20	175.01	124.65	71.36	97.50
	33	52.18	140.59	179.36	78.88	96.88
Almost unsupervised	2	52.19	176.38	153.37	91.42	97.50
	3	15.23	155.32	145.85	63.20	99.38
	5	34.97	126.98	149.77	64.91	98.75
	9	15.23	168.19	147.59	66.24	99.38
	23	97.97	168.38	176.41	125.60	92.50
	24	105.09	146.89	205.22	123.40	90.00
	25	150.37	169.19	216.97	150.16	82.50
	28	69.82	166.42	164.11	93.33	95.63
	32	53.37	142.10	185.44	99.94	96.88
	36	81.49	154.82	193.75	116.22	94.38
	37	104.82	160.20	200.55	92.31	90.00
	39	51.68	162.09	176.85	87.80	96.88
	40	59.34	166.72	174.26	102.01	96.88

The best fitting model is underlined

DIM-O unidimensional classifier based on the orientation, *DIM-F* unidimensional classifier based on the frequency, *GCC* General Conjunctive Classifier, *GLC* General Linear Classifier

condition, 11 participants of the 13 learned successfully: they revealed a decision bound based on the general linear classifier. Participants 25 and 37 preferred a decision bound based on the general conjunctive classifier, indicating that they switched to another strategy in comparison to day 2.

Figures 9 and 10 in the Supplementary Materials show the actual responses during the last two blocks of day 4 for each participant. Table 9 presents the model fits on the test day (i.e., day 5). Again, for the almost supervised condition all participants used a decision bound based on the general linear classifier. In the almost unsupervised condition, 11 participants preferred a decision bound based on the general linear classifier whereas participants 23 and 37 adopted a strategy based on the general conjunctive classifier. Figures 11 and 12 in the Supplementary Materials present the responses and the best fitting decision bounds for every participant during the test day.

Analysis of the response times The mean response times (RTs) along with the 95% confidence intervals from days 1–4 for the almost supervised, almost unsupervised, and the semisupervised condition of Experiment 1 from days 1–4 are presented in Fig. 7. These RTs were calculated on the last two blocks of each day. For day 1, the mean RT in the almost supervised condition was 855 ms (*SD* =162.41), the mean RT in the almost unsupervised condition was 825 ms (*SD*=156.31), and the mean RT in the semisupervised condition of Experiment was 858 ms (*SD*=131.22). Importantly for this investigation is that participants were equally fast in the last two blocks of day 2 (almost supervised mean RT=806 ms, *SD*=161.29; almost unsupervised mean RT=800 ms, *SD*=148.49, and semisupervised mean RT=785 ms, *SD*=90.10). This was confirmed by a one-way ANOVA comparing the RTs of the last two blocks of day 2 to the three conditions, almost supervised condition, the almost unsupervised condition, and the semisupervised condition of Experiment 1, F <1, p=.93, η_p^2=.004. On days 3 and 4, the mean RTs dropped in the almost supervised condition to, respectively, 783 ms (*SD*=170.67) and 772 ms (*SD*=134.68). This decrease in mean RTs was also observed in the almost unsupervised condition: 772 ms (*SD*=134.68) on day 3 and 749 ms (*SD*=144.54) on day 4. A repeated measures ANOVA

Table 9 Mean accuracy (%) and model-based analyses (BIC scores) on all trials of day 5 for every participant of Experiment 2

Condition	Participant	GLC	DIM-O	DIM-F	GCC	Mean accuracy (%)
Almost supervised	6	63.81	181.73	155.45	104.78	92.53
	8	57.49	171.93	165.23	109.59	95.53
	11	98.93	170.44	178.41	121.78	89.56
	12	93.64	176.87	171.30	115.55	90.29
	14	59.40	178.45	159.58	95.90	94.77
	15	137.33	191.39	174.85	145.64	81.35
	17	85.57	166.05	179.45	88.06	90.29
	18	65.02	154.51	184.79	111.94	93.29
	19	58.48	171.18	165.67	109.62	93.29
	20	25.86	165.86	165.83	101.39	98.50
	21	71.43	162.14	179.59	122.45	94.02
	22	58.08	173.09	164.69	109.56	94.77
	30	69.64	164.96	175.69	115.15	94.02
	31	78.10	136.74	193.03	99.51	88.80
	33	88.22	151.80	193.36	125.25	88.05
Almost unsupervised	2	74.23	222.84	170.29	120.24	94.77
	3	87.93	233.26	184.67	126.47	89.55
	5	102.91	226.44	196.58	127.34	82.83
	9	112.09	250.90	198.29	119.71	76.11
	23	146.69	216.83	180.00	130.53	79.86
	24	66.29	225.13	191.32	123.43	91.80
	25	102.65	229.95	211.97	119.91	81.35
	28	81.91	226.28	201.71	102.30	85.83
	32	103.47	222.37	193.09	123.79	87.32
	36	116.88	222.70	187.54	125.02	85.83
	37	112.96	234.52	216.60	79.68	76.86
	39	71.56	219.76	151.07	97.00	91.05
	40	49.29	223.11	166.02	106.68	95.53

The best fitting model is underlined

DIM-O unidimensional classifier based on the orientation, *DIM-F* unidimensional classifier based on the frequency, *GCC* General Conjunctive Classifier, *GLC* General Linear Classifier

was conducted to determine whether the mean RTs on the last two blocks differed depending on the day (four levels: days 1, 2, 3, and 4) and condition (three levels: almost supervised, almost unsupervised, and semisupervised of Experiment 1). Not surprisingly, there was main effect of day, $F(3,34)=6.81$, $p=.001$, $\eta_p^2=.38$, indicating that the RTs decreased across the succeeding days. Post hoc paired-sample *t*-tests, adjusted by the Bonferroni correction for multiple comparisons, indicated that participants were significantly faster on days 2, 3, and 4 compared to day 1, $t(38)=2.92$, $p=.03$; $t(38)=2.95$, $p=.05$ and $t(38)=4.03$, $p=.001$, respectively. None of the other comparisons were significant (all p>.42). There was no main effect of condition, $F<1$, $p=.89$, $\eta_p^2=.01$, neither was there a significant interaction between day and condition, $F(6,68)=1.85$, $p=.10$, $\eta_p^2=.14$.

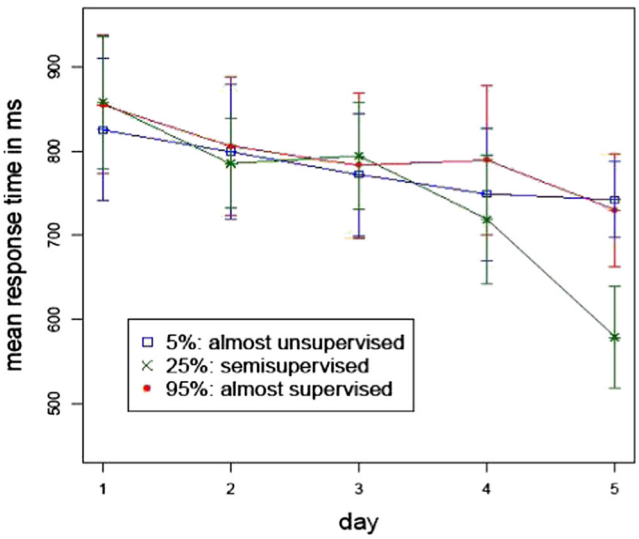

Fig. 7 Mean response times (in ms) along with the 95% confidence intervals for the almost supervised and the almost unsupervised condition of Experiment 2 along with the semisupervised condition of Experiment 1. The mean response times are calculated on the last two blocks of each day

Decisive to test our hypotheses was the difference in RTs on day 5. In the almost supervised condition the mean RT was 730 ms (*SD*=*132.08*), the mean RT in the almost unsupervised condition was 742 ms (*SD* =84.59), and the mean RT in the semisupervised condition of Experiment 1 was 579 ms (*SD*=104.08). A one-way ANOVA indicated significant differences between the RTs of the three groups on day 5, $F(2,36)=8.07$, $p=.001$, $\eta_p^2=.31$. Post hoc paired-sample *t*-tests, adjusted by the Bonferroni correction for multiple comparisons, showed that RTs were faster in the semisupervised condition of Experiment 1, compared to the almost supervised condition, $t(24)=3.15$, $p=.004$, and to the almost unsupervised

condition, $t(22)=4.26$, $p=.003$. The difference between the almost supervised condition and the almost unsupervised condition was not significant, $t(26)=0.28$, $p>.99$.

Speed-accuracy trade-off On day 5 participants were deliberately asked to respond as fast as possible. The speed-accuracy trade-off was again calculated for each condition. In the almost supervised condition, there was no SAT-effect, $r=-.08$, $p=.79$. In contrast, there was a SAT-effect in the almost unsupervised condition, $r=.85$, $p<.001$: faster responses were correlated with more errors. Recall that in the semisupervised condition of Experiment 1 we also observed a SAT-effect.

Discussion

In Experiment 2 the total number of trials was equal in each condition to investigate whether the nature of feedback (almost supervised, semisupervised, or almost unsupervised) had an impact on the development of automaticity. Two control conditions were run to compare the semisupervised condition of Experiment 1: the almost supervised condition (in which mainly feedback trials were given) and the almost unsupervised condition (in which mainly no-feedback trials were given). The results indicated that the mean RTs on day 5 in the semisupervised condition of Experiment 1 were significantly faster than in the two control conditions of Experiment 2, indicating that a combination of feedback and no-feedback trials boosted the automaticity process. This result cannot be due to general faster RTs in the semisupervised condition, as RTs were equal for all conditions on day 2. This result can also not be affected by the novelty of the no-feedback trials on the test day, since all participants already experienced feedback and no-feedback trials on days 3 and 4. Remarkably, in the almost unsupervised condition, two participants did not reveal a strategy based upon the optimal decision bound anymore on day 4. Apparently, they changed their categorization strategy during day 3 and day 4. This was not the case in the almost supervised condition nor in the semisupervised condition. This might imply that for some participants a minimal percentage of feedback is still needed late in learning – even though almost perfect accuracy was obtained long before. On the test day, the mean accuracy was higher in the almost supervised condition compared to the almost unsupervised condition. There was no significant difference in mean accuracy in the semisupervised condition compared to the almost supervised condition and the almost unsupervised condition. In the semisupervised and the almost unsupervised condition, a few participants showed a drop in accuracy on day 5. This drop in performance was also reflected in the model-based analyses: in the almost supervised condition all participants adhered to a strategy based upon the optimal decision bound whereas in the semisupervised and the almost unsupervised conditions three and two participants, respectively, switched. When these participants who switched strategy were omitted from the analyses, a one-way ANOVA revealed that the difference in accuracy disappeared, $F(2,31)=2.67$, $p=.12$ $\eta_p^2=.13$. The mean accuracy was 87% *(SD=5.90)* for the almost unsupervised, 89% *(SD*=7.04*)* for the semisupervised, and 92% *(SD=4.08)* for the almost supervised condition. Still, the effect of faster RTs in the semisupervised condition remained, one-way ANOVA $F(2,31)=6.54$, $p=.004$, $\eta_p^2=.30$. Furthermore, this effect was due to the significant difference between the semisupervised (577ms, *SD*=123.14) and the almost supervised (730 ms, *SD*=132.08), $t(21)=2.71$, $p=.015$ and the difference between the semisupervised and the almost unsupervised (760 ms, *SD*=79.20), $t(17)=3.96$, $p=.005$ while the difference between the almost supervised and the almost unsupervised condition was not significant, $t(24)=0.67$, $p>.99$. These post hoc *t*-tests were corrected for multiple comparisons by Bonferroni. These results show that the faster response times in the semisupervised condition were not due to quick random guessing by some participants, resulting in fast RTs and low accuracy, since the effect remains when the switch participants were omitted.

On day 5 a SAT-effect occurred in the almost unsupervised condition of Experiment 2: participants who tended to respond fast, also made more errors. This was not the case in the almost supervised condition. This could explain the significantly lower accuracy in the almost unsupervised condition on day 5. Note that we also observed a SAT-effect on day 5 for the semisupervised condition. Even though this condition showed a SAT, it has a comparable accuracy as the almost supervised condition.

As a conclusion, Experiment 2 shows that, even when the total number of trials is the same, the development of automaticity is enhanced by 25% semisupervised learning. Only in this feedback scheme the accuracy remained high and response times were strikingly faster. When feedback was almost always provided, participants maintained a high accuracy but were not able to accelerate their responses. In the almost unsupervised condition, there is a drop in accuracy – some participants even unlearned the category structure – and response times remained high. These results are in contrast to the SPEED-model, which predicts a similar development of automaticity between the three conditions as the number of trials, regardless of feedback, is identical in all conditions.

General discussion

This study investigated the impact of semisupervised category learning late in the learning process when automaticity develops. Participants were first trained in a supervised way over 2 days on the information-integration category structure and only participants who performed at least 90% accurately and

used a decision bound similar to the optimal decision bound were included in the actual experiments. In Experiment 1, half of the participants were trained in a 25% semisupervised way on days 3 and 4: only a quarter of the trials were followed by feedback. The other half were trained in a supervised way, implying that feedback was given after every categorization response. Both conditions received an equal amount of feedback trials. On the fifth day, differences in performance between the semisupervised and supervised learners were studied. Participants were urged to respond as fast as possible on this test day. Accuracy was similar in both groups on day 5, which is to be expected, as accuracy was already above 90% at the end of day 2. However, the results clearly showed that participants in the semisupervised condition responded significantly faster than the participants in the supervised condition on day 5. This effect cannot be due to general faster RTs and/or higher accuracy levels in the semisupervised condition, as evidenced by similar RTs and accuracies for both conditions at the end of day 2. Thus, the findings of Experiment 1 imply that late in learning the no-feedback trials in the semisupervised condition aided the development of automaticity.

However, two confounds hampered a clear conclusion that late in learning semisupervised learning is superior. First, even though the amount of feedback trials was equal in the semisupervised and the supervised conditions, the total number of trials differed. Participants in the semisupervised condition received four times as many trials on day 3 and day 4 as participants in the supervised condition and hence had more practice. It is therefore possible that the larger total number of trials caused the faster RTs in the semisupervised condition. Second, participants in the supervised condition never received no-feedback trials on days 3 and 4. Perhaps the sudden encounter of no-feedback trials on day 5 might have slowed them down on this test day. To exclude these alternative explanations, two control conditions were administered in Experiment 2 that were compared to the semisupervised condition of Experiment 1. The percentage of feedback trials on day 3 and day 4 was manipulated. In the almost supervised condition 95% of the trials were randomly followed by feedback. In the almost unsupervised condition 5% of the trials were randomly followed by feedback. This implies that in all conditions (almost supervised, semisupervised, and almost unsupervised), participants encountered both feedback and no-feedback trials on days 3 and 4. Crucially, the total number of trials was identical in all three conditions. Despite these alterations, the results of Experiment 2 again showed that the RTs on the fifth day were significantly slower for the almost supervised and the almost unsupervised conditions compared to the semisupervised condition, whereas RTs did not differ at the end of day 2. Since the total number of trials was now identical in all three conditions, we can conclude that the semisupervised learners achieved automaticity faster. Hence, late in learning semisupervised learning should be preferred. These results are not in line with the predictions of the SPEED-model since it stipulates that the type of trial (feedback or no feedback) should not have an impact on the development of automaticity late in learning, and would therefore expect similar RTs in all the conditions as they all contain the same number of trials. Why semisupervised category learning is superior late in learning requires further investigation. Perhaps semisupervised learning increased participants' motivation and attention compared to a condition where feedback is almost always offered. Contrarily, participants in the almost unsupervised learning condition often reported frustration and they perhaps gave up to perform to their maximum ability.

In the almost unsupervised condition there was a Speed accuracy trade-off (SAT-)effect: participants who tended to respond fast also made more errors. This can explain why the mean accuracy in the almost supervised condition was significantly higher than in the almost unsupervised condition. Although we also observed a SAT-effect in the semisupervised condition, there was no difference in accuracy between the almost supervised and the semisupervised conditions, suggesting that, even though participants in the semisupervised condition sacrificed accuracy for speed, they still remained at a similar accuracy level to participants in the almost supervised condition. The model-based analyses on day 5 showed that a few participants in the semisupervised and the almost unsupervised condition switched in strategy. This was not the case in the almost supervised condition where all participants adhered to the optimal decision bound. Apparently, some participants in the almost unsupervised and in the semisupervised condition unlearned the category structure. Note that when these participants were omitted from analyses the faster RTs in the semisupervised condition remained, whereas the difference in accuracy between almost supervised and almost unsupervised conditions disappeared. These findings suggest that the faster RTs in the semisupervised condition were not due to quick random guessing. It rather seems that individual differences are decisive. In this study, the learning scheme in a condition was fixed and individual differences in learning were not taken into account. It is possible that some participants need a higher percentage of feedback-trials on days 3 and 4 even though they successfully learned the structure at the end of day 2. Another explanation could be that the switch point, that is, the point in the learning process on which supervised learning becomes less effective in favor of semisupervised learning, is later for some participants than the fixed point of 800 trials (end of day 2) used in this study. Thus, even though semisupervised learning appears to be the best learning mode late in learning, there may be exceptions for some participants. These individual differences in learning are interesting directions for further research.

The current study supports the idea that no-feedback trials aid the learning process, as shown in machine learning

(Chapelle et al., 2006; Zhu et al., 2009) and in a few human semisupervised studies (Kalish et al., 2011; Lake & McClelland, 2011; Zhu et al., 2010). Contrary to the studies of Kalish et al. (2011), Lake and McClelland (2011) and Zhu et al. (2010), in our study all the stimuli (as compared to a fixed subset) could be followed by feedback, mimicking real-life category learning. As in the study of Rogers et al. (2010), semisupervised learning was found when the participants were urged to respond as fast as possible. It is possible that, in order to observe semisupervised learning, speed of responding is essential. Again, this can be an interesting direction for future research. Our study also proves that not only young children (Kalish et al., 2015), but young adults too are able to learn in a semisupervised way.

The results of our study may seem to be in contrast to the study of Vandist et al. (2009) where no effect of the no-feedback trials was found in learning the information-integration category structure. However, there are important differences between both studies. First and most importantly, Vandist et al. (2009) focused on the effects early in the learning process whereas the current study dealt with the effects late in learning, when automaticity develops. Second, the impact of the no-feedback trials was studied on the accuracy levels in the study of Vandist et al. (2009). In this study response time was the dependent variable. Third, in the Vandist et al. (2009) study, participants in the semisupervised condition learned from the start in a semisupervised way whereas in the present study semisupervised learning was only introduced after expert performance was obtained.

Combining the results of both studies suggests that early in learning the information-integration structure, the no-feedback trials do not have an impact, but that late in learning the no-feedback trials facilitate automaticity. Indeed, the effects of semisupervised learning, at least in the information-integration structure, might be especially apparent late in the learning process. These results can also explain why former semisupervised category studies failed to find convincing effects, as they all focused on initial learning processes. This also makes sense if we relate this to category learning in children. When a child is first confronted with items of an unknown category, parents label most of the presented items. As the child becomes more and more familiar with the category, the parent still labels information but less often. When the parent has the idea that the child has acquired the category, label information diminishes but still takes place from time to time. In fact, it is ecologically plausible that semisupervised learning takes place when a solid basis of category expertise has first been acquired and from that point on, it aids learning. Experiment 2 even suggests that from a certain expert level on, semisupervised category learning might be essential to the development of automaticity, since semisupervised category learning fastens the development of automaticity whereas (almost) supervised learning seems to slow it down. Although speculative, it could for example be that the continuous feedback in the almost supervised condition makes the expert learner less attentive or less motivated, leading to decreased performance. Nevertheless, our results also indicated that a certain percentage of feedback trials is still needed to develop automaticity successfully: when feedback was rare in the almost unsupervised condition, the mean RTs remained at the same level and at the end of day 4 a few participants even unlearned the category structure.

In conclusion, this is the first study that examines the effect of 25% semisupervised learning late in learning. In Experiment 1, faster RTs were observed in a 25% semisupervised condition in comparison to a supervised condition when the total amount of feedback trials in both conditions was the same. In Experiment 2 the total number of trials was kept identical in all conditions, but still the 25% semisupervised condition of Experiment 1 showed faster response times in comparison to the almost supervised and almost unsupervised conditions of Experiment 2. Hence, late in learning (25%), semisupervised learning seemed to have a beneficial effect as the no-feedback trials facilitated automaticity. A learning condition containing a certain amount of no-feedback trials even seems to outperform a condition where (almost) all trials are followed by feedback, as long as a minimal percentage of the trials is followed by feedback.

Appendix

Four models were fit to each participant's data: two rule-based decision bound models, i.e., the vertical and the horizontal unidimensional models, and one information-integration model, i.e., the general linear classifier. More details of the models can be found in Ashby & Gott (1988), Ashby (1992), and Maddox and Ashby (1993).

Rule-based models

The unidimensional classifier (DIM). The DIM assumes that participants set a decision criterion on one of the stimulus dimensions. Because our stimuli have two stimulus dimensions, orientation and frequency, two unidimensional classifiers were fit: one for the frequency dimension and one for the orientation dimension. An example of a unidimensional rule used for categorization is: "Respond A if the orientation is steep, otherwise respond B." In this case the frequency is irrelevant. These models have two parameters: a decision criterion along the chosen classification dimension and a perceptual noise variance parameter.

The general conjunctive classifier (GCC). The GCC assumes that the rule used by participants is a conjunction, for

example: "Respond A if the orientation is steep and the frequency is small, otherwise respond B." The GCC has three parameters: one for each for the single-decision criterion placed along each dimension (i.e., orientation and frequency) and finally one for the perceptual noise variance.

Information-integration model

The general linear classifier (GLC). The GLC assumes that a linear decision bound divided the stimulus space into response regions. Confronted with a stimulus, the perceived place in the stimulus space is determined and the contributed categorization response elicited. These decision bounds require linear integration of both stimulus dimensions, resulting in an information-integration decision strategy. The GLC has three parameters: the slope and the intercept of the linear decision bound and finally the perceptual noise variance.

References

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale: Erlbaum.

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

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences,* **5**,204-210.

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, *31*(7), 1114–1125.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*, 632–656.

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*, 33–53.

Ashby, F. G., & Maddox, W. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37* (3), 372-400.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178.

Ashby, F. G., & Maddox, W. T. (2010). Human category learning 2.0. *Annals of the New York Academy of Sciences, 1224,* 147-161.

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, *30*, 666–677.

Ashby, F. G., & O'Brien, J. B. (2007). The effects of positive versus negative feedback on information-integration category learning. *Perception & Psychophysics*, *69*, 865–878.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*, 1178–1199.

Ashby, F. G., Turner, B. O., & Horvitz, J. C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Sciences, 14*(5), 191-232.

Ashby, F. G., & Crossley, M. (2012). Automaticity and multiple memory systems. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*, 353–376.

Censor, N., Karni, A., & Sagi, D. (2006). A link between perceptual learning, adaptation and sleep, *Vision Research*, *46*, 4071–4074.

Chapelle, O., Schölkopf, B., Zien, A (2006). Semi-supervised learning. MIT Press, Cambridge, MA, USA.

Clapper, J. P., & Bower, G. H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *20*, 443–460.

Ell, S. W., & Ashby, F. G. (2006). The effects of category overlap on information-integration and rule-based category learning. *Perception & Psychophysics*, *68*, 1013–1026.

French, R. M., Mareschal, D., Mermillod, M., & Quinn, P. C. (2004). The role of bottom-up processing in perceptual categorization by 3- to 4-month-old infants: simulations and data. *Journal of experimental psychology: General, 133* (3), 382-397.

Gibson, B. R., Rogers, T. T., & Zhu, X. (2013). Human semi-supervised learning. *Topics in Cognitive Science*, *5*, 132–172.

Gibson, B.R., Rogers, T.T., Kalish, C.W., & Zhu, X (2015). What causes categoryshifting in human semi-supervised learning? In Proceedings of the 37th Annual Conference of the Cognitive Science Society (CogSci)

Heitz, R. P. (2014). The speed-accuracy trade-off: history, physiology, methodology and behavior. *Frontiers in Neuroscience, 8* (150), 1-15.

Hélie, S., Waldschmidt, J. G., & Ashby, F. G. (2010). Automaticity in rule-based and information-integration categorization. *Attention, Perception & Psychophysics*, *72*, 1013–1031.

Kalish, C. W., Rogers, T. T., Lang, J., & Zhu, X. (2011). Can semi-supervised learning explain incorrect beliefs about categories? *Cognition*, *120*, 106–118.

Kalish, C.W., Zhu, X., & Rogers, T.T. (2015). Drift in children's categories: when experienced distributions conflict with prior learning. *Developmental Science. 18(6),* 940-956.

Lake, B. M., & McClelland, J. L. (2011). Estimating the strength of unlabeled information during semi-supervised learning. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 1400–1405.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492-527.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*, 829–835.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*, 49–70.

Maddox, W. T., Ashby, F. G., & Gottlob, L. R. (1998). Response time distributions in multidimensional perceptual categorization. *Perception & Psychophysics*, *60*(4), 620–637.

Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004a). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & cognition, 32* (4), 582-591.

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004b). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review*, *11*, 945–952.

Maddox, W. T., & Filoteo, J. V. (2011). Stimulus range and discontinuity effects on information-integration category learning and generalization. *Attention, Perception & Psychophysics*, *73*, 1279–1295.

Maddox, W. T., Filoteo, J. V., Hejl, K. D., & Ing, A. D. (2004c). Category number impacts rule-based but not information-integration category learning: further evidence for dissociable category-learning systems. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*, 227–245.

Maddox, W. T., Glass, B. D., O'Brien, J. B., Filoteo, J. V., & Ashby, F. G. (2010a). Category label and response location shifts in category learning. *Psychological Research*, *74*, 219–236.

Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system

in perceptual category learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*, 100–107.

Maddox, W. T., Pacheco, J., Reeves, M., Zhu, B., & Schnyer, D. M. (2010b). Rule-based and information-integration category learning in normal aging. *Neuropsychologia*, *48*, 2998–3008.

McDonnell, J.V., Jew, C.J., and Gureckis, T.M. (2012). Sparse category labels obstruct generalization of category membership In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology. Human Perception and Performance*, *21*, 128–148.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85* (3), 207-238.

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

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

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

Moors, A., & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. *Psychological Bulletin*, *132*, 297–326.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *13*, 87–108.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104,* 266-300.

Paul, E. J., Boomer, J., Smith, J. D., & Ashby, F. G. (2011). Information-integration category learning and the human uncertainty response. *Memory & Cognition*, *39*, 536–554.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26 (3),* 303-343.

Pothos, E. M., & Chater, N. (2005). Unsupervised categorization and category learning. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *58*, 733–752.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition*, *121*, 83–100.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skilss. *Journal of Experimental Psychology: General, 126,* 288-311.

Rogers, T. T., Kalish, C., Gibson, B. R., Harrison, J., & Zhu, X. (2010). Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2320–2325).

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: Behavior, theory, and biological mechanisms, *Cognitive Science, 27* (3), 525-559.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Shepard, R., Hovland, C., & Jenkins, H. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*, 1–42.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84* (2), 127-190.

Spiering, B. J., & Ashby, F. G. (2008a). Response processes in information-integration category learning. *Neurobiology of Learning and Memory*, *90*, 330–338.

Spiering, B. J., & Ashby, F. G. (2008b). Initial training with difficult items facilitates information-integration, but not rule-based category learning: Research article. *Psychological Science*, *19*, 1169–1177.

Stevens, M., Lammertyn, J., Verbruggen, F., & Vandierendonck, A. (2006). Tscope: A C library for programming cognitive experiments on the MS windows platform. *Behavior Research Methods*, *38*, 280–286.

Stickgold, R., & Walker, M. P. (2005). Memory consolidation and reconsolidation: What is the role of sleep? Trends in Neuroscience, 28, 408–415

Stickgold, R., James, L., & Hobson, J. A. (2000a). Visual discrimination learning requires sleep after training. *Nature Neuroscience*, *3*, 1237–1238.

Stickgold, R., Whidbee, D., Schirmer, B., Patel, V., & Hobson, J. A. (2000b). Visual discrimination task improvement: A multi-step process occurring during sleep. *Journal of Cognitive Neuroscience*, *12*, 246–254.

Vandist, K., De Schryver, M., & Rosseel, Y. (2009). Semisupervised category learning: the impact of feedback in learning the information-integration task. *Attention, Perception & Psychophysics*, *71*(2), 328–341.

Vermaercke, B., Cop, E., Willems, S., D'Hooge, R., & Op de Beeck, H. P. (2014). More complex brains are not always better: rats outperform humans in implicit category-based generalization by implementing a similarity-based strategy. *Psychonomic Bulletin & Review*, *21*, 1080–6.

Vong, W. K., Perfors, A., & Navarro, D. J. (2014). The Relevance of Labels in Semi-Supervised Learning Depends on Category Structure. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society,* 1718–1723.

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

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

Zeithamova, D., & Maddox, W. T. (2007). The role of visuospatial and verbal working memory in perceptual category learning. *Memory & Cognition*, *35*, 1380–1398.

Zhu, X., Gibson, B.R., Jun, K.-S., Rogers, T.T., Harrison, J. & Kalish, C. (2010). Cognitive models of test-item effects in human category learning. In *Proceedings of the 27th International Conference on Machine Learning,* 1247–1254.

Zhu, X., & Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. Morgan and Claypool Publishers