Artif Intell Rev
DOI 10.1007/s10462-007-9024-7

Using knowledge partitioning to investigate the psychological plausibility of mixtures of experts

Sébastien Hélie · Gyslain Giguère · Denis Cousineau ·
Robert Proulx

Abstract Over the years, the presence of knowledge partitioning (KP) in human function learning data has been used to argue that mixture-of-experts models (MOE) constitute a psychologically plausible explanation of human performance, and that the experts used by humans are always linear. These claims recently led to the proposition of the population of linear experts model (POLE). In this paper, variations of the firefighting paradigm developed by Lewandowsky and his colleagues, which initiated research about KP, were used to explore the psychological plausibility of MOE in general and POLE in particular. In a first experiment, these statements were tested by modifying the test display of the firefighting paradigm. The results showed that adding irrelevant information to the display resulted in a smaller proportion of partitioning participants. Also, some participants used non-linear experts to partition the stimulus space. This new type of KP was further explored in a second study, which included more training sessions. The results suggest that linear KP disappears with practice and that non-linear partitioning reflects the incapacity to correctly estimate the position of the function's vertex. It is concluded that MOE are adequate psychological models, but that the linearity and ubiquity claims of the POLE model need to be weakened.

Keywords Cognitive model · Function learning · Knowledge partitioning · Mixture-of-experts · Psychology

S. Hélie (✉)
Rensselaer Polytechnic Institute, Troy, NY, USA
e-mail: helies@rpi.edu

G. Giguère · R. Proulx
Université du Québec À Montréal, Montreal, QC, Canada
e-mail: giguere.gyslain@courrier.uqam.ca

R. Proulx
e-mail: proulx.robert@uqam.ca

D. Cousineau
Université de Montréal, Montreal, QC, Canada
e-mail: denis.cousineau@umontreal.ca

1 Introduction

The nature of intelligence has fascinated researchers in the fields of philosophy, psychology, and computer science for many years. In his classic 1986 book, Marvin Minsky made a strong statement about the complexity of human intelligence: the mind is far too complex to be the result of a single, homogenous, general principle (Minsky 1986). Hence, general intelligence must be the result of aggregating several specialized agents, which function in qualitatively different manners, and communicate in order to achieve the objectives of the "creature" (O'Leary 2005). This hypothesis was called the "Society of Minds" (Minsky 1986). In a society of minds, there is no centralized control and the resulting behaviour surpasses by far the capacities of the individual agents.

Because Minsky's book only provided a high-level description of his theory, and left the specific implementation details for future work, many different models have been proposed in the past 20 years. Chief among the tentative implementations is the blackboard system (Corkill 1991; Epstein 1992, 1994), a particularly effective small-scale society of minds (Singh 2003). This implementation is based on the following metaphor: imagine several specialists from different fields, surrounding a standard blackboard on which a problem has been written. Each specialist may try to solve a part of the problem and write his piece on the blackboard, which can be used by other specialists to further solve the problem at hand. This process continues until the problem is completely solved. Among the characteristics of a blackboard system listed by Corkill (1991), those that are particularly relevant for this discussion are listed here. First, the specialists are independent: that is, each one is fully functional in the absence of the others (modularity). Second, because the specialists are modular, their internal representations and inferencing machinery are hidden from their counterparts, and they can vary from one specialist to the next. As a result, any information can be put on the blackboard because different specialists use different representations (e.g., decision trees versus production rules). However, there must be a common interaction language (interface) which ensures that what is put on the blackboard is understood by every specialist. This communication constraint is strengthened by the fact that no expert is able to individually solve the entire problem: the generation of the solution is incremental, and each step constitutes a small contribution to the overall solution. As a result, there is a need for a control unit, which directs traffic and decides which expert is allowed to write on the blackboard at any given moment. For instance, if the problem to be solved is of an algebraic form, the mathematician should be allowed to write on the blackboard before the psychologist.

While blackboard architectures have been used in several application domains (e.g., game playing, case-based reasoning, sensory interpretation, data fusion, etc.), they suffer from two main shortcomings: scalability and communication (Singh 2003). Both these deficiencies are related: first, when there are a few hundred specialists, the design of an interface that allows each specialist to understand what the others are writing on the blackboard becomes a quasi-impossible task. Second, when an interface is successfully conceived, a large number of independent experts usually produce many conflicting solutions, which can be simultaneously present on the blackboard, and there is no well-understood solution to settle the conflicts. One way to avoid these problems is to prevent, instead of require, communication between the specialists. This is what mixtures of experts do.

1.1 Mixtures of experts

A mixture-of-experts (MOE) model is an architecture composed of a gating mechanism and several expert modules (Bishop 1995; Jacobs et al. 1991). When using such a model, the problem is fed to the gating mechanism, which classifies it in order to dispatch the input to the correct expert. Once an expert receives the problem, it computes a complete solution. Hence, a MOE architecture is akin to a blackboard system except that each expert (specialist) is able to completely solve some of the problems; it is the gating mechanism's (control unit) role to know which expert can solve a given problem. Therefore, the experts need not communicate directly. This class of models is particularly effective in situations where different (even contradictory) responses are appropriate according to situations (Bishop 1995), such as multispeaker vowel recognition (Jacobs et al. 1991).

1.2 Function learning and its application to forest fires

Just like the gating mechanisms of mixtures of experts, humans must learn to classify everyday situations according to context in order to choose the correct action to undertake. In the particular case where exemplars (inputs) and categories (outputs) are continuous, one usually extracts the function relating the input to the output (function learning), instead of performing standard associative learning (DeLosh et al. 1997). Simple examples of function learning include estimating the distance of a moving object according to its size on the retina, or how long you can stay in the sun before you burn (Harris and Minda 2005).

Another situation where learning a function is necessary is when estimating the speed of spread of forest fires (Lewandowsky et al. 2002; Lewandowsky and Kirsner 2000). When the slope of the terrain and the wind direction are in opposition, a forest fire spreads uphill at a speed negatively related to wind speed, unless the wind becomes strong enough to overcome the fire's natural propensity to spread uphill. From this moment on, the fire spreads downhill at a speed positively related to the increasing wind speed. Overall, the function relating speed of spread to wind speed is a concave quadratic function where the vertex indicates the point at which the force applied by the wind overpowers the tendency of the fire to spread uphill (see Fig. 3a for an example).

Another important aspect of firefighting is the use of back-burning fires to control the reach of the to-be-controlled fire. Back-burning fires are lit and managed by firefighters to starve the to-be-controlled fire of fuel. Usually, a back-burner is used when the wind speed is low; otherwise the firefighters might lose control of this second fire.

The type of fire (back-burning, to-be-controlled) is an important cue which facilitates the estimation of a fire's spreading speed: instead of considering a quadratic function to determine the propagation of the fire, firefighters can use two linear functions: a decreasing function associated to the back-burning context and an increasing one associated to the to-be-controlled context. This two-stage decision process (cue identification followed by response selection) is equivalent to a mixture-of-experts architecture, which identifies the context first and uses a different linear expert accordingly. In Lewandowsky and Kirsner's paper (2000), experienced firefighters were shown to use this two-stage strategy. The authors argued that the association between the context (type of fire) and the linear functions had been learnt through their many years of experience. By extension, it was argued that this two-stage process was ubiquitous in expertise. This theory was called knowledge partitioning (KP).

To test the KP theory, another experiment was designed to assess whether novices would also use this strategy in a function learning task (Lewandowsky et al. 2002). In this second experiment, the participants were taught basic firefighting background knowledge and trained in a standard function learning experiment using a concave quadratic function. Every stimulus (wind speed) was also accompanied by a context label, which was systematically associated to a different half of the function during training (Fig. 1). This manipulation aimed at recreating the bias present in experienced firefighters' knowledge, for which back-burners are

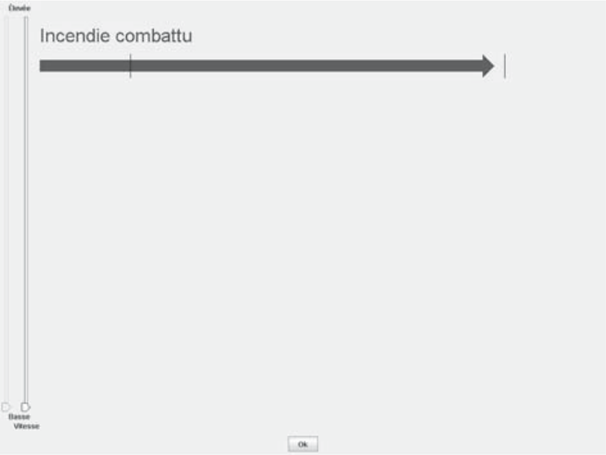

Fig. 1 Example display of the firefighting paradigm used to study knowledge partitioning (screenshot from the extra information condition, Experiment 1). In the other conditions, the display was identical, except for the absence of visual markers (vertical lines across the stimulus)

usually encountered in low wind speed situations and to-be-controlled fires in high wind speed conditions. At test, every stimulus was presented twice: once as a back-burner and once as a to-be-controlled fire. The results showed that participants easily achieved the task. In particular, spreading speeds were almost perfectly estimated when wind speeds appeared in their usual context. However, they were systematically underestimated when shown in the unusual context. Hence, participants gave different responses to identical stimuli presented in different contexts, which supports the KP theory.

These results constitute empirical evidence in favour of the psychological plausibility of mixture-of-experts models (Heit and Bott 2000; Little et al. in press). In particular, one such model was proposed to explain human performance: population of linear experts (POLE: Kalish et al. 2004). In POLE, when a stimulus is encountered, a gating mechanism directs it to the correct expert, which represents one of many linear functions with different slopes and intercepts. There are enough experts to cover the entire stimulus space and only the gating system has adjustable weights. Once an expert is chosen, it computes the answer accordingly.

This model possesses three important properties. First, POLE always uses KP and can explain all past results in the function learning literature. Second, experts do not blend together. Therefore, the system always commits to a cue-value and chooses an expert accordingly (KP). Third, each expert represents a linear relationship between the stimulus and the response. These properties, and the preceding empirical results, were used to make certain claims concerning the generality and properties of KP (Kalish et al. 2004; Lewandowsky et al. 2002): (1) KP is always used when the association between the context and a part of the function is systematic, (2) the usefulness of KP is proportional to task difficulty, and (3) humans always partition into *linear* sub-functions to achieve complex function learning tasks. In the following experiments, we empirically tested the preceding claims related to KP in general and to POLE in particular.

In Experiment 1, human data was collected by altering Lewandowsky et al.'s (2002) experimental settings. A first group performed the same task as Lewandowsky et al.'s Experiment 1, systematic-context condition. A second group was trained in the same task with stimuli covering a smaller part of the domain: this task was hypothesized to be more difficult and should lead to an equal or higher proportion of partitioners. The third group was trained with settings identical to the second, except that irrelevant information was added to the display (constant visual markers). Because the necessary conditions for finding KP were present in this condition (e.g., a systematic association between particular values and a context), we should find as many partitioners as in the second group.

In Experiment 2, the genesis of KP was explored by studying its relation with expertise more thoroughly. Hence, the task was made perceptually more difficult by using variations on the most difficult condition of Experiment 1 (group 2), and the training period was extended. As a result, it was possible to compute the proportion of participants using KP at the beginning of training and compare it to the proportion after extensive training. Also, because a new type of partitioning was found in the second group of Experiment 1, its presence was assumed to be related to task difficulty. Hence, the increased difficulty of Experiment 2 allowed verification of this hypothesis.

2 Experiment 1

This experiment is an extension of Lewandowsky et al.'s (2002) Experiment 1, systematic-context condition. Therefore, this section bears on their original methodology. It was brought to our attention that our participants were given more extensive background knowledge than in Lewandowsky et al.'s original study (M. Kalish, personal communication). Nevertheless, performance was not qualitatively different, as shown by the results from the control group.

2.1 Experiment 1: method

2.1.1 Participants

Fifty-four undergraduate students from the Université de Montréal participated in this experiment. The control group was composed of eighteen participants, which were trained in a reproduction of Lewandowsky et al.'s (2002) experiment. Another 18 participants were trained with stimuli covering a smaller part of the display (small stimuli group), and the remaining participants were trained with small stimuli and supplemental information (extra information group). In each group, six participants were assigned to the complete condition, six to the left-only condition, and the remaining six to the right-only condition (the difference between these conditions are described next). Participants in the complete conditions received 7$ as compensation for their time, and those in the left-only or right-only conditions received 5$. The experiment was conducted in French.

2.1.2 Material

Each participant was tested individually. All instructions and stimuli were presented on 43 cm (17 in.) monitors connected to PCs, and the participants were positioned approximately 60 cm away from the monitor. The experimental task was programmed using Sun Microsystems' Java J2SDK1.4.1. The program was used to present the material and record the participants' answers.

2.1.3 Stimuli

The participants were expected to learn a concave quadratic function in which the fire's spreading speed (F) was related to wind speed (W) in the following manner: $F(W) = 24.2 - 1.8W + 0.05W^2$. Wind direction always opposed slope, and the vertex of the function $(W = 18)$ represented the point at which the force of the wind balanced the effect of the slope. To the left of that point $(W < 18)$, fire speed decreased with increasing wind speed, without changing the direction of the fire spread. Lewandowsky et al. (2002) referred to these fires as "slope-driven". To the right of the vertex $(W > 18)$, fires were "wind-driven" and their speed increased as a function of wind speed. During training, 36 stimuli were used, ranging from wind speeds of 0 to 36, omitting the vertex of the function. At test, the omitted wind speed of 18 was included, resulting in a total of 37 transfer stimuli.

On each trial, a horizontal arrow, whose length was proportional to a particular wind speed (henceforth referred to as the stimulus), was shown at the top of the display. The minimal arrow length, associated with the value 0, was approximately 0.7 cm for the control group and 5.8 cm for the small stimuli and extra information groups. The maximal length, associated with the value 36, was approximately 31 cm for the control group and 26 cm for the small stimuli and extra information groups. Thus, in the small stimuli and the extra information groups, the shortest arrow occupied 1/6 of the display and the longest 5/6. In the control group, the arrows spanned the entire display. No numerical values for wind or fire speed were shown. Participants were to consider each fire in a context represented both by a color-coded arrow and textual label (blue for *Back-burning* and red for *Firefighting*). In the extra information group, visual markers were added to the display to indicate the minimal and maximal possible stimulus lengths (two small vertical bars); they were the only difference between the small stimuli group and the extra information group. Because the markers were constant, they could not be used by the hypothesized gating mechanism and thus, the results in this condition should be identical to those obtained in the small stimuli condition (see Fig. 1 for an example of a trial in the extra information condition).

The participants were asked to predict the speed of the fire (notwithstanding its direction of spread) by moving a vertical sliding pointer along a 23 cm scale positioned in the left part of the display. The scale was labelled *slow* at the bottom and *fast* at the top, without any incremental values or tick marks. After each training trial, the participant's response was followed by the addition of a feedback arrow. The arrow was located next to the response scale to indicate the correct speed of spread. Also, a message appeared in a rectangle at the bottom center of the screen to encourage the participant to perform better (yellow rectangle) or to indicate that the response was satisfying (green rectangle). Predictions deviating by five or more units (approximately 7.2 cm) from the correct answer were accompanied by the former (yellow message) while acceptable performances were accompanied by the latter (green message). Participants were required to acknowledge feedback by a mouse click. The inter-stimulus interval (ISI) was 2 s and the textual context-label always preceded the stimulus by 1 s. At test, feedback was absent.

2.1.4 Procedure

The procedure was identical for all groups and varied across conditions. In the complete condition, there was a total of 180 training trials, 90% of which included stimuli appearing in their associated context (wind-driven fires were associated to the firefighting context while slope-driven fires were associated to the back-burning context); the remaining 10% of the trials included stimuli appearing in the other context. In the left-only and right-only conditions, the training was restricted to one-half of the function ($W < 18$ for the left-only condition and $W > 18$ for the right-only), resulting in 90 training trials. In these conditions, all stimuli were presented in their associated context.

In all conditions, every stimulus length was presented once within each block of 36 trials (18 for the left-only and the right-only conditions), except during the first block of the complete condition, in which all stimuli associated with a given context were first presented, followed by all the stimuli associated with the other context.

After completion of the training trials, participants in all conditions completed the same transfer test. The transfer test involved predicting the fire speed of all stimuli in both contexts.

2.2 Experiment 1: results

2.2.1 Training

The participants' Absolute Deviation from Function (ADF) was used to evaluate learning. The performance of participant 222 (from the small stimuli group, complete condition), deteriorated with practice ($F(4, 175) = 2.54$, $p < 0.05$). Therefore, this participant was excluded from the following analyses.

The learning curves are shown in Fig. 2. As seen, participants in all conditions from all groups improved their ADF and were thus able to learn the function. Also, Fig. 2 suggests no effects of groups or conditions.

A Group (small stimuli versus extra information versus control) × Condition (complete, left-only, right-only) × Block (5, repeated measures) ANOVA was performed on the participants' ADF to corroborate what Fig. 2 hinted. First, the participants were able to reduce their ADF with practice: The mean ADF was 2.33 in the first block and diminished to 1.74 in the fifth ($F(4, 176) = 14.32$, $p < 0.01$). However, this effect must be interpreted with care, because the Block × Group interaction was significant ($F(8, 176) = 10.24$, $p < 0.01$). The group effect was further decomposed within each block. The groups significantly differed in the first block of training ($F(2, 44) = 28.42$, $p < 0.01$) but were similar in all other blocks (all $F(2, 44) < 1.63$, $p > 0.05$). *Tukey A post hoc comparisons* showed that the control group was significantly better than the other two at the beginning of the task (both differences > 0.96, $p < 0.01$). However, as suggested by the absence of group effect in the remaining blocks, this difference disappeared with training.

2.2.2 Group performance at test

KP can be detected experimentally by a difference in responses to a given stimulus shown in different contexts (Lewandowsky et al. 2002). Figure 3 shows transfer performances for participants trained in the complete conditions. As seen in the top-left panel, participants trained with extra information learned the function quite well. Surprisingly, answers in both contexts matched the quadratic function perfectly: hence, there was no effect of context.

Responses of participants trained with small stimuli are shown in the top-right panel. As expected, their responses at test were affected by the context (compare with the top-left panel), but in an unexpected way. In comparison, the deviations found by Lewandowsky et al. (2002) were systematic: low wind speeds resulted in an underestimation of the speeds of fire spread in the firefighting context and high wind speeds accordingly resulted in underestimations in the back-burning context. This is exactly the pattern of results found in the control group (see the bottom-left panel). In the top-right panel, the underestimations are present (to a lesser extent), but mid-range wind speeds were overestimated.

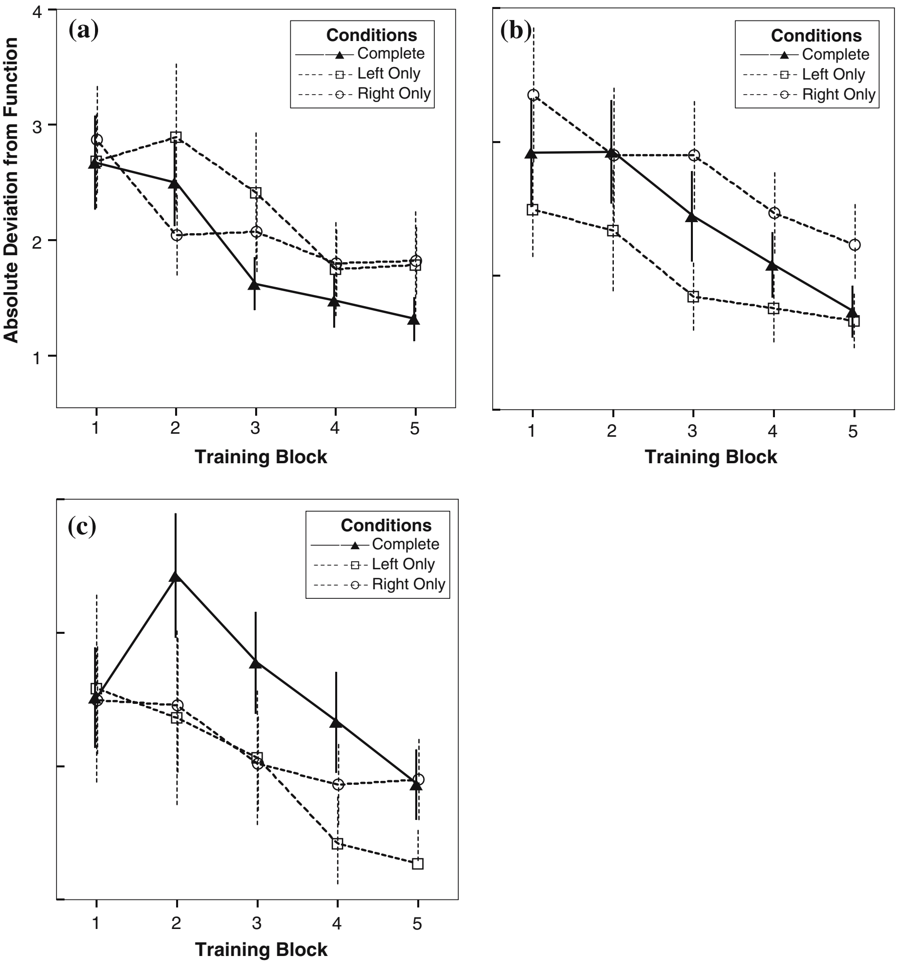

Fig. 2 Participants' ADF during the training phase. The top-left panel (**a**) shows the performance of participants in the extra information group, the top-right panel (**b**) shows the performance of participants in the small stimuli group, and the bottom-left panel (**c**) shows the control group

A better way to highlight the difference in responses to a given stimulus is to compute the signed differences (Lewandowsky et al. 2002). A signed difference is computed by subtracting the answer given at test to each stimulus in the back-burning context from the answer given to the same stimulus in the firefighting context. Signed differences randomly aggregated around zero would suggest the absence of partitioning, while signed differences systematically deviating in one direction would indicate the presence of partitioning.

The top-left panel of Fig. 4 confirms that participants trained with extra information are not partitioning their knowledge: their signed differences are randomly aggregated around the x-axis (RMSD = 2.69). Participants in the control group did show the expected pattern of results (bottom-left panel): signed differences are negative to the left of the vertex and positive to the right (RMSD = 5.50). The signed differences of participants trained with small

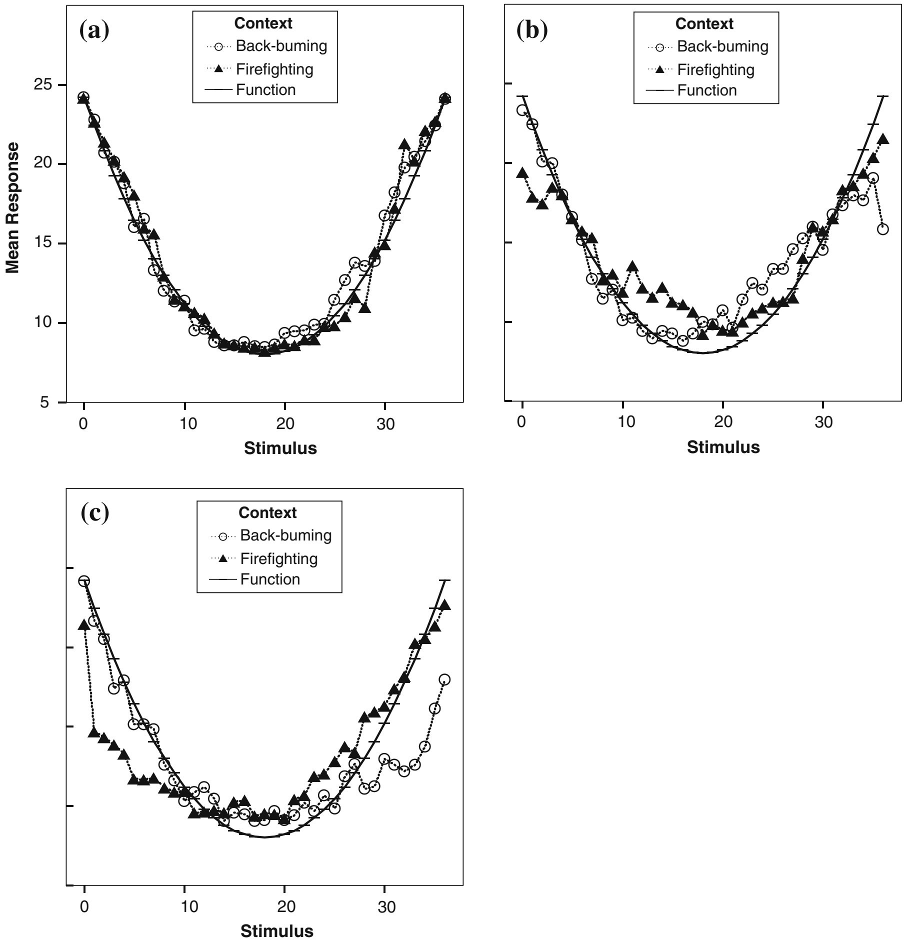

Fig. 3 Mean responses at test for each context. Panels represent the same groups as in Fig. 2

stimuli are more intriguing (top-right panel): they are substantially deviating from the x-axis in a sine-like manner (RMSD = 5.27).

2.2.3 Individual results at test

Considering that Lewandowsky et al. (2002) found important individual differences relating to KP, it is relevant to verify if the effects found in the previous section were representative of the entire groups of participants. A novel statistical way of classifying the participants as partitioners (P) or non-partitioners (NP) is to individually plot their signed differences and estimate the best-fitting linear model using a linear regression. A slope, which is significantly different from zero suggests a systematic effect of context, namely KP. On the other hand, a slope of zero suggests no clear effect of context. Table 1 shows the slope and intercept individually estimated for each participant.

As seen in Table 1, all but one participant fit a model with an absolute slope of 0.05 or less in the group trained with additional information. Because these slopes did not differ

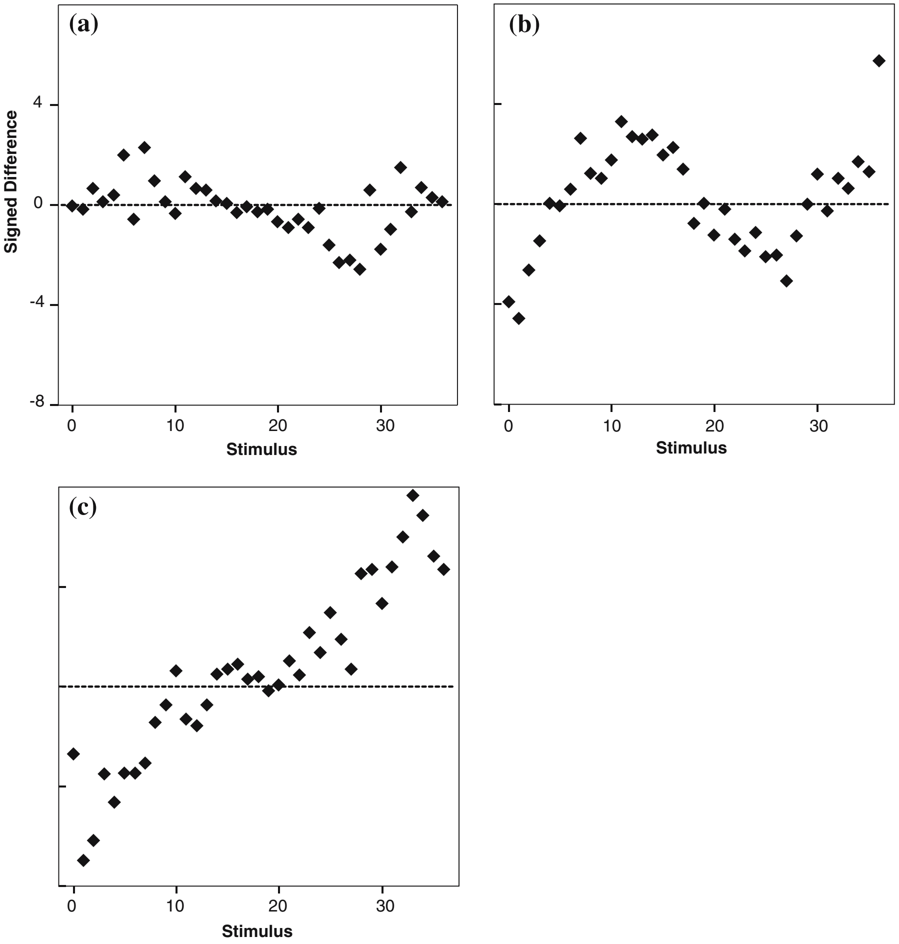

Fig. 4 Mean signed differences at test. Panels represent the same groups as in Fig. 2

significantly from zero ($p > 0.05$), these participants were classified as NPs. The exact opposite was true of participants in the small stimuli group: all but one participant had an absolute slope greater than 0.20. Hence, these participants were classified as Ps. In the control group, the best-fitting slopes of two of the six participants were smaller than 0.10: these slopes did not significantly differ from zero ($p > 0.05$) and these participants were classified as NPs. The four remaining participants were classified as Ps.

The proportion of Ps in the small stimuli group significantly differed from the proportion of Ps in the extra information group according to a binomial test ($B(5, 1/6) = 4, p < 0.01$). The proportion of Ps in the small stimuli group (80%) is well in range with past literature[1] while the proportion of Ps in the group trained with extra information (16.7%) is below past results. The proportion of Ps in the control group (67%) is similar to Lewandowsky et al.'s (2002) results and does not significantly differ from the small stimuli

[1] To be exact, previous research found between 13% (Yang and Lewandosky 2003) and 50% (Kalish et al. 2004) of participants who were not partitioning their knowledge.

Table 1 Estimated parameters for the best-fitting linear models

	Participants	Estimated parameters			
		Slope	Intercept	r^2	Classification
Extra information					
	110	−0.03	0.63	0.02	NP
	111	0.00	−0.13	0.00	NP
	112	−0.25	5.48	0.34	P
	120	−0.02	0.37	0.01	NP
	121	0.05	−1.45	0.07	NP
	122	0.01	−1.51	0.00	NP
Small stimuli					
	210	−0.41	8.03	0.58	P
	211	−0.02	0.19	0.00	NP
	212	0.22	−4.19	0.29	P
	220	0.63	−9.04	0.93	P
	221	−0.23	2.68	0.32	P
Control					
	310	0.60	−10.74	0.87	P
	311	−0.02	0.17	0.01	NP
	312	−0.07	2.16	0.09	NP
	320	0.54	−9.45	0.69	P
	321	0.13	−2.61	0.14	P
	322	0.65	−8.88	0.89	P

Note: P, Partitioners; NP, Non-partitioners

group ($B(6, 4/5) = 4$, $p > 0.05$). However, this proportion of Ps differs from the proportion found in the extra information group ($B(6, 1/6) = 4$, $p < 0.01$).

Together, these results suggest that when extra information is present in the display, fewer participants use the KP heuristic, even if the added information is less useful than the context. Also, it is noteworthy that all the Ps in the control group showed positive slopes, which is consistent with the "linear experts" hypothesis (Kalish et al. 2004; Lewandowsky et al. 2002). However, half of the Ps in the small stimuli group and the only P in the extra information group had negative slopes, which is consistent with the sine-like pattern of Fig. 4b. The overestimation of moderate wind speeds is also present in the top-right panel of Fig. 3 and further inspection suggests a partitioning of the stimuli in two quadratic functions with skewed vertices. Therefore, diminishing the stimulus' length does not prevent participants from using KP but entails a different, non-linear, type of partitioning, which is not consistent with POLE's predictions.

2.2.4 Independence of knowledge parcels

As Lewandowsky et al. (2002) pointed out, participants who were uniquely trained on the left or right part of the function represent extreme cases of KP: they possess a single expert, associated with a single context. Therefore, if the knowledge of Ps in each context is truly independent, their responses should be similar to the left-only condition in the back-burning context and the right-only condition in the firefighting context.

In the case of non-linear Ps, responses in the back-burning context were similar to responses from participants uniquely trained in this particular context (left-only condition: $r = 0.87$). However, the correlation between partitioners' responses in the firefighting context and those from the right-only condition was smaller ($r = 0.69$). This difference is significant using the Fisher Z transform test ($Z = 2.0$, $p < 0.05$). Therefore, the back-burning parcel of knowledge seems more hermetic than the firefighting parcel. Also, results from Lewandowsky et al. suggested higher correlation coefficients.

In the case of linear Ps, responses from knowledge partitioners were similar to responses from participants trained in the left-only ($r = 0.81$) and right-only ($r = 0.83$) conditions (in the back-burning and firefighting contexts, respectively). Also, the difference between correlation coefficients is not statistically significant ($Z = 0.25$, $p > 0.05$). Knowledge about the other half of the function acquired in another context did not affect the participants' responses, suggesting that knowledge was completely partitioned.

2.3 Experiment 1: discussion

In Experiment 1, the usual settings used to assess the presence of KP (Kalish et al. 2004; Lewandowsky et al. 2002) were varied to check the robustness of this phenomenon. Precisely, two modifications were made: reducing the length of the stimuli, and adding potentially distracting information. First, results from the control group confirmed the adequacy of our reproduction of Lewandowsky et al.'s experiment. Second, it is well established that diminishing the span of the stimuli increases discrimination difficulty (Goldstein 1999), hence making stimulus estimation more difficult. In the small stimuli group, the participants used KP in an expected proportion, but showed negatively sloped best-fitting linear models (Table 1). This counter-intuitive result was first hinted by sine-like signed differences (Fig. 4b) and the use of non-linear expert functions with skewed vertices (Fig. 3b). Finally, adding useless information to the display was sufficient to prevent most participants from using KP to achieve the task, even if the sufficient conditions for the use of this strategy were present. However, those participants who did not use KP to simplify the function were still able to learn it (as shown by an absence of group effect in the ANOVA).

Following these results, there is little doubt that KP is the heuristic of choice when no other information is available and participants notice a systematic association between a particular context and a class of stimuli. However, the detection of non-linear KP is a substantive contribution of Experiment 1, which needs to be further explored. For instance, Fig. 3b suggested that non-linear partitioners were using two quadratic experts with skewed vertices. Hence, non-linear KP might be the result of a difficulty to correctly estimate the vertex of the function, the point at which the association between the stimuli and a particular context ends. Experiment 2 explored this hypothesis in two ways. First, the difficulty of estimation of the vertex was varied by altering the position of the arrow on the display. When a stimulus is not centered in its frame, its center is more difficult to estimate (anchor effect: Goldstein 1999). Hence, the small stimuli condition was replicated, along with two other conditions in which the stimuli were of the exact same length, but either anchored to the left or the right of the display. Second, because the most difficult condition of Experiment 1 is assumed to be the easiest of Experiment 2, some participants were trained for three sessions, instead of one. Therefore, if non-linear partitioning is related to the difficulty to estimate the vertex of the function, it should disappear after extensive training.

3 Experiment 2

3.1 Experiment 2: method

3.1.1 Participants

Thirty-six undergraduate students from the Université de Montréal participated in this experiment, twelve of which were trained in the *anchored left* condition (AL), another twelve in the *anchored right* condition (AR), and the remaining in the *centered* condition (a reproduction of the small stimuli condition of Experiment 1; henceforth referred to as C). In each condition, half the participants were trained for a single session and the remaining participated in three sessions of training. The participants received 7$ and 20$, respectively as a compensation for their time.

3.1.2 Material

The material was identical to that of Experiment 1.

3.1.3 Stimuli

In the C condition, the stimuli used were identical to those in Experiment 1, small stimuli condition. In the AL condition, the stimuli were identical to those in the C condition (same scale), except that the arrow was anchored to the left of the display. Hence, the smallest stimulus was 0.7 cm long and the longest was 20.9 cm long. In the AR condition, the scale of the stimuli was identical to the other two conditions, but the stimuli were anchored to the right. Hence, the smallest stimulus was 10.8 cm long and the longest was 31 cm long.

3.1.4 Procedure

The procedure for the participants trained in a single session was identical to that of Experiment 1. For participants trained during three sessions, each session was identical to the one described in Experiment 1; hence, there were three transfer tests (one after each session). The sessions took place on three consecutive working days.

3.2 Experiment 2: results

3.2.1 Training

As in Experiment 1, the training performance was assessed by using the ADF. The performance of participant 1524 (from the C condition) worsened with practice: ($F(4, 175) = 4.76$, $p < 0.01$); this participant was eliminated from the following analysis. Also, the data from participant 3312 (in the AL condition) was lost due to computer problems.

The learning curves are shown in Fig. 5. Separate Condition (AL, AR, C) × Block (5 or 15, repeated measures) ANOVAs were performed on the training performance of the participants who participated in one (panel a) or three (panel b) training sessions. For the participants trained in a single session, the Block main effect was significant ($F(4, 56) = 8.92$, $p < 0.01$); the mean ADF in the first block was 3.63 and diminished to 2.75 at the end of training. There are no between condition differences ($F(2, 24) = 0.81$, $p > 0.05$), but the Block × Condition interaction reached significance ($F(8, 56) = 3.42$, $p < 0.05$). *Tukey A post hoc*

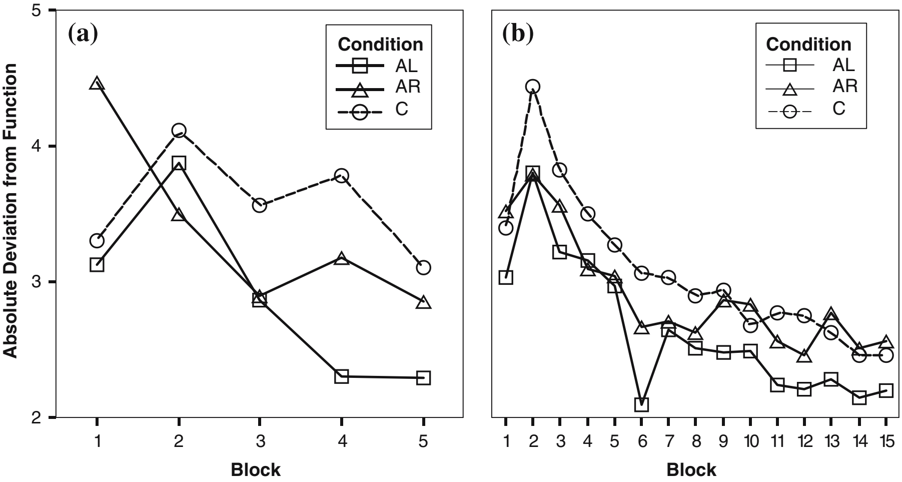

Fig. 5 Participants' ADF during the training phase. The left panel (**a**) shows the performance of participants trained in a single session while the right panel (**b**) shows the performance of participants trained for three sessions

comparisons showed that the participants in the AL condition were significantly better than those in the AR condition during the first block of training ($p < 0.05$). Also, the participants in the AL conditions were significantly better than those in the C condition during block four ($p < 0.05$). There was no other effect of Condition.

The ANOVA on the ADF of the participants trained for three sessions showed similar results. The participants' ADF was 3.31 at the beginning of training and diminished to 2.41 at the end of the third session. This difference is significant ($F(14, 196) = 9.44$, $p < 0.01$). Similar to the participants trained in a single session, there was no effect of the Condition factor ($F(2, 14) = 0.23$, $p > 0.05$). However, the Block × Condition interaction failed to reach significance in this case ($F(28, 196) = 0.32$, $p > 0.05$).

Overall, participants in all conditions were able to learn the function and there were no reliable differences between the conditions, notwithstanding the length of the training phase. Because the participants achieved the task, it is now possible to individually classify them according to their learning strategy.

3.2.2 Classification of the performance

One of the focuses of this experiment was the linearity of the parcels of knowledge; hence, a more elaborate technique was used to classify the participants. Because the non-linear partitioners of Experiment 1 had signed differences similar to one cycle of a sine function, cubic linear models were individually fit to the participants' signed differences. A cubic or quadratic coefficient significantly differing from zero suggested non-linear partitioning (P-NL). When only the linear coefficient significantly differed from zero, the participant was classified as a linear partitioner (P-L), and when none of the coefficients significantly differed from zero[2], the participant was classified as a non-partitioner (NP). Reclassification of the participants in Experiment 1 using this new criterion yielded the same classification result, thus confirming the validity of this statistical model. The classification results for the first training session are

[2] Because the intercept did not provide relevant information, it was ignored in the present analysis.

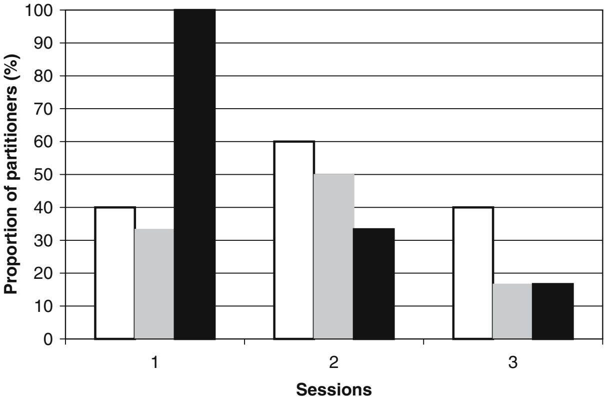

Fig. 6 Proportion of participants using KP in each session (participants trained for three sessions only). The black bars represent the C condition, the white bars represent the AL condition, and the grey bars represent the AR condition

shown in Table 2. Overall, 19 participants (56%) used the KP strategy during the first session of training. Among the partitioners, 15 had non-linear parcels of knowledge (79%) while the remaining used linear experts.

Cubic linear models were also fit to the signed differences in the second and third tests of participants who attended to three experimental sessions (see Table 3). The proportion of participants using KP was reduced to 47% in the second session and diminished to 14% in the third! These proportions needed to be split by group, to observe whether this trend reflected each of them equally well.

Figure 6 shows the proportion of partitioners in each condition for each session (notwithstanding the type of partitioning). Surprisingly, the proportion of partitioners in the AL and AR conditions does not depend on the number of practice sessions: the same proportion of participants used this heuristic in all three sessions (both absolute slopes of the best fitting linear model = 8.35, $p > 0.05$). However, the situation is quite different for the C condition: there is a negative linear trend in the proportion of participants using the KP heuristic (absolute slope of the best fitting linear model = 41.7, $p < 0.05$). As a result, in the first session, there are far more partitioners in the C condition than in the other two. During the other sessions, the between-condition differences are negligible. The reliability of these differences was corroborated by confidence intervals: if the number of partitioners in each session is independent from the condition (i.e., each participant has the same chance of being a partitioner, ignoring condition), the number of partitioners in each session is binomially distributed, with parameters n = number of partitioners and p = 1 / number of conditions = 1/3. Hence, the first session is B(10, 1/3), the second session is B(8, 1/3), and the last is B(4, 1/3). As a result, the 95% bicaudal confidence intervals are [1, 6], [0, 5], and [0, 3], respectively for sessions one to three. The C condition in the first session lies on the edge of the confidence interval, which suggests a deviation from the assumed distribution. Besides this effect, all conditions in all sessions did not significantly deviate from their postulated distribution.

Contrary to our previous hypothesis, the participants did not seem to evolve from non-linear partitioning to linear partitioning. Participants in the C condition went from partitioning to non-partitioning, and there was no effect of practice on the proportion of partitioners in the other conditions. To further explore this effect, the participants were split by partitioning type and the result for each condition is shown in Fig. 7. As seen, there are always more

Table 2 Estimated best-fitting cubic model during the first training session

		Estimated				
	Participant	x^3	x^2	x	r^2	Classification
AL						
	1310	8.34	−13.31	4.74	0.58	P-NL
	1311	−0.98	−0.19	1.45	0.34	NP
	1312	0.47	−0.41	0.52	0.86	NP
	1320	4.47	−7.31	2.43	0.40	P-NL
	1321	4.23	−7.46	3.16	0.16	P-NL
	1322	3.08	−4.43	1.38	0.06	NP
	3311	−1.41	2.37	−0.10	0.83	NP
	3313	4.09	−7.33	3.62	0.21	P-NL
	3321	3.66	−5.41	1.75	0.09	NP
	3322	−0.50	−1.12	1.41	0.25	NP
	3323	3.50	−6.01	3.51	0.88	P-NL
AR						
	1410	4.32	−5.92	1.17	0.50	P-NL
	1411	4.70	−7.01	3.04	0.44	P-NL
	1412	1.91	−3.03	1.40	0.06	NP
	1420	−0.12	1.12	−1.46	0.29	NP
	1421	3.41	−4.83	0.86	0.55	P-NL
	1422	−4.57	6.78	−2.19	0.14	NP
	3411	3.77	−5.33	2.28	0.42	NP
	3412	−0.41	0.86	0.36	0.03	NP
	3413	5.10	−8.03	3.78	0.58	P-NL
	3421	1.61	−2.31	0.67	0.02	NP
	3422	0.36	−0.08	0.73	0.97	P-L
	3423	3.35	−5.00	2.29	0.33	NP
C						
	1514	4.60	3.10	0.57	0.57	P-NL
	1515	2.29	1.08	0.12	0.12	NP
	1516	2.20	3.44	0.46	0.46	P-L
	1525	0.03	−0.44	0.11	0.11	NP
	1526	0.40	0.28	0.03	0.03	NP
	3511	4.75	5.22	0.63	0.63	P-NL
	3512	0.21	1.60	0.95	0.95	P-L
	3513	5.16	2.21	0.38	0.38	P-NL
	3521	3.12	3.14	0.96	0.96	P-NL
	3522	−4.53	−1.81	0.57	0.57	P-NL
	3523	0.16	−2.46	0.52	0.52	P-L

Note: P-L, Linear partitioning; P-NL, Non-linear partitioning; NP, No partitioning

Table 3 Estimated best-fitting cubic model during the second and third training sessions

Participant	Estimated x^3	x^2	x	r^2	Classification
AL (Session 2)					
3311	0.00	0.01	−0.09	0.02	NP
3313	0.00	−0.09	2.10	0.84	P-NL
3321	0.00	−0.01	0.15	0.18	NP
3322	0.00	−0.08	1.02	0.66	P-NL
3323	0.00	−0.15	2.78	0.77	P-NL
AL (Session 3)					
3311	0.00	−0.05	0.62	0.12	NP
3313	0.00	−0.14	2.67	0.78	P-NL
3321	0.00	0.00	0.01	0.02	NP
3322	0.00	−0.48	0.57	0.43	P-NL
3323	0.00	0.01	−0.26	0.02	NP
AR (Session 2)					
3411	0.00	−0.05	0.68	0.61	P-NL
3412	0.00	−0.04	0.52	0.16	NP
3413	0.00	−0.01	0.39	0.05	NP
3421	0.00	0.02	−0.26	0.03	NP
3422	0.00	−0.01	0.98	0.95	P-L
3423	0.00	0.02	−0.41	0.24	P-NL
AR (Session 3)					
3411	0.00	−0.12	2.44	0.53	P-NL
3412	0.00	0.08	−1.21	0.11	NP
3413	0.00	0.03	−0.39	0.03	NP
3421	0.00	−0.01	0.16	0.06	NP
3422	0.00	0.04	−0.77	0.75	NP
3423	0.00	0.00	−0.20	0.22	NP
C (Session 2)					
3511	0.00	0.00	0.00	0.01	NP
3512	0.00	−0.09	2.36	0.93	P-NL
3513	0.00	0.01	−0.08	0.00	NP
3521	0.00	0.02	0.05	0.15	NP
3522	0.00	0.03	−0.36	0.72	P-NL
3523	0.00	−0.01	0.17	0.08	NP
C (Session 3)					
3511	0.00	0.02	−0.44	0.11	NP
3512	0.00	0.04	−0.54	0.03	NP
3513	0.00	−0.02	0.27	0.05	NP
3521	0.00	0.06	−0.59	0.24	P-NL
3522	0.00	0.01	−0.10	0.57	NP
3523	0.00	−0.04	0.52	0.18	NP

Note: P-L, Linear partitioning; P-NL, Non-linear partitioning; NP, No partitioning

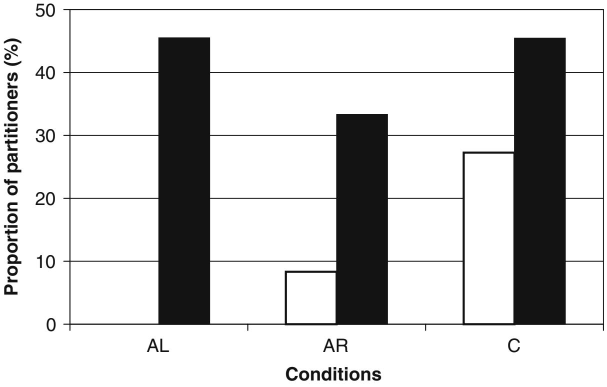

Fig. 7 Proportion of partitioning participants by type (linear versus non-linear) for each group (in their first training session). The white bars represent the proportion of linear partitioners while the black bars represent the proportion of non-linear partitioners. In each group, the addition of the white and black bars yields the proportion of partitioners

non-linear than linear partitioners. Also, the larger number of partitioners in the C condition during the first session (Fig. 6) seems to be the result of a larger number of linear partitioners in this condition: in each condition, we always find about the same proportion of non-linear partitioners (∼40%) but there are more linear partitioners in the C condition than in the others (∼30%). A closer look at the individual classification (Table 3) shows that it is the linear partitioners who changed strategy and stopped using KP: in the second session of practice, only one participant used linear partitioning and in the third, no participants used linear experts.

3.3 Experiment 2: discussion

The aim of Experiment 2 was twofold: (1) explore the distinction between linear and non-linear partitioning and, (2) explore the genesis of KP. Concerning the first topic, the results supported our previous hypothesis: non-linear partitioning is present when the participants are trying to use the contextual information but are unable to correctly identify the vertex of the function. This was suggested by the (quasi-) absence of linear partitioners in the AR and AL conditions. However, the C condition included both linear and non-linear partitioners, and the proportions were the same as in the small stimuli condition of Experiment 1, thus confirming the stability of the observed phenomenon.

Concerning the genesis of partitioning, the results were quite interesting. Linear partitioners seemed to abandon this strategy after extensive training while non-linear partitioners kept using this heuristic. The former result should not surprise the reader. Lewandowsky and his colleagues argued that linear partitioning is a suboptimal strategy to achieve the task (Kalish et al. 2004; Lewandowsky et al. 2002; Lewandowsky and Kirsner 2000): it results in estimation errors when the stimuli are encountered in their less frequent context (whereas ignoring the context and learning the function results in perfect estimation). In the knowledge restructuring literature, it was shown that two conditions are necessary to elicit a change in strategy (Kalish et al. 2005; Little et al. in press): (1) the participants must make mistakes and, (2) another strategy must be readily available. In the present paper, it was clearly shown that linear partitioning led to important underestimation in the infrequent context and learning the function was a clear alternative. However, Experiment 1 showed that non-linear partitioning leads to smaller estimation error (Fig. 3b), which might be insufficient to encourage the participants to switch strategy.

4 Summary and conclusions

In the preceding experiments, it was shown that: (1) the presence of supplemental information can prevent participants from using KP, (2) two types of KP (linear and non-linear) can be observed, (3) the proportion of participants using linear KP decreases with practice, and (4) the proportion of participants using linear KP is negatively related to the difficulty to estimate the function's vertex. These results have important implications for both empirical researches related to KP and psychological modelling.

4.1 Implications for current research on knowledge partitioning

In the KP experimental paradigm (Lewandowsky et al. 2002), the most salient information is the context: it appears on the screen before the rest of the display and it is identified by more than one source (textual label, display color). Hence, participants are trying to use this information whenever possible. In order to be able to use contextual information, there must be a systematic link between the context and a part of the function. Also, participants must be able to correctly identify the portion of the function associated to each context (i.e., identify the vertex of the function). When both these conditions are met, linear partitioning can result. However, this heuristic is suboptimal and is abandoned when the function is correctly learned (after extensive training). If the participants are unable to identify the portion of the function associated to each contextual value, the contextual information is useless and KP will not be used; the function is directly learned. However, some participants might be unable to correctly estimate the position of the vertex but still try to use KP. This behaviour results in non-linear partitioning, which might provide a *good enough* estimation of the function; hence, this strategy persists. However, because this hypothesis partly results from an exploratory research with few participants, further research, including more participants to increase statistical power, is needed to assess its validity.

4.2 Implications for current cognitive modelling

The present results constitute empirical evidence in favour of the psychological plausibility of mixture-of-experts models (Bishop 1995; Jacobs et al. 1991). When contextual information is systematically linked to a particular task, human participants seem to be using a two-stage decision process to simplify the functional relationship between the current state of the world and the action to be undertaken: first, the context is identified, then the correct action is chosen. However, in order to efficiently use this strategy, the participants must be able to correctly infer the gating mechanism. The present data suggests that this can be achieved when no other information interferes with the identification of the context and the range of the functional relationship is known.

The collected data also has important implications for the psychological plausibility of the POLE model (Kalish et al. 2004). This model was created to explain human function learning and categorization in the presence of systematic contextual information. Apparently, POLE's postulates seem to underestimate the capacity of humans to learn non-linear mapping. In the present experiments, non-linear experts were successfully used to learn a quadratic function, which is contrary to one of POLE's most basic assumptions: the exclusive presence of linear experts. Nevertheless, non-linear experts were not the only type of experts detected: the data also suggested the presence of linear experts, which is consistent with POLE's assumptions. Hence, as proposed by the Society of minds hypothesis (Minsky 1986), and its blackboard implementation (Corkill 1991; Epstein 1992, 1994), heterogeneous agents seem to be responsible for human intelligence and psychological modellers should reconsider their search for a single, homogeneous, mechanism responsible for human intelligence.

Acknowledgements The authors would like to thank Dominic Charbonneau and Sonja Engmann for their help with the participants in the experiments, and two anonymous reviewers for their useful comments on an earlier version of this manuscript. This research was supported by scholarships and grants from *Le fonds québecois de la recherche sur la nature et les technologies, Le fonds québecois de la recherche sur la société et les cultures*, and the *Natural Sciences and Engineering Research Council of Canada*.

References

1. Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, New York
2. Corkill DD (1991) Blackboard systems. AI Experts 6:40–47
3. DeLosh EL, Busemeyer JR, McDaniel MA (1997) Extrapolation: the sine qua non for abstraction in function learning. J Exp Psychol Learn Mem Cogn 23:968–986
4. Epstein SL (1992) The role of memory and concepts in learning. Minds Machines 2:239–265
5. Epstein SL (1994) For the right reasons: the FORR architecture for learning in a skill domain. Cogn Sci 18:479–511
6. Goldstein EB (1999) Sensation & perception, 5th edn. Brooks/Cole Publishing Company: Pacific Grove, CA.
7. Harris HD, Minda JP (2005) Function learning with an ensemble of linear experts and off-the-shelf category-learning models. In: Bara BG, Barsalou L, Bucciarelli M (eds) Proceedings of the 27th annual conference of the cognitive science society, Erlbaum Associates, Mahwah, NJ, pp 905–910
8. Heit E, Bott L (2000) Knowledge selection in category learning. Psychol Learn Motiv 39:163–199
9. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. Neural Comput 3:79–87
10. Kalish ML, Lewandowsky S, Davies M (2005) Error-driven knowledge restructuring in categorization. J Exp Psychol Learn Mem Cogn 31:846–861
11. Kalish ML, Lewandowsky S, Krushcke JK (2004) Population of linear experts: Knowledge partitioning and function learning. Psychol Rev 111:1072–1099
12. Lewandowsky S, Kalish M, Ngang SK (2002) Simplified learning in complex situations: Knowledge partitioning in function learning. J Exp Psychol Gen 131:163–193
13. Lewandowsky S, Kirsner K (2000) Knowledge partitioning: context-dependent use of expertise. Mem Cogn 28:295–305
14. Little DR, Lewandowsky S, Heit E (in press) Ad hoc category restructuring. Mem Cogn
15. Minsky M (1986) The society of mind. Simon and Schuster, New York
16. O'Leary C (2005) Reuse and arbitration in diverse societies of mind. In: Proceedings of The sixteenth Irish conference on artificial intelligence and cognitive science, Portstewart, University of Ulster, pp 369–378
17. Singh P (2003) Examining the society of mind. Comput Informat 22:521–543
18. Yang L-X, Lewandowsky S (2003) Context-gated knowledge partitioning in categorization. J Exp Psychol Learn Mem Cogn 29:663–679