Learning and Individual Differences 32 (2014) 19–25

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

Learning to judge creativity: The underlying mechanisms in creativity training for non-expert judges

Martin Storme [a,b,*], Nils Myszkowski [a,b], Pinar Çelik [a], Todd Lubart [a]

[a] Laboratoire Adaptations Travail Individu, Université Paris Descartes, 71 Avenue Edouard Vaillant, Boulogne-Billancourt, France
[b] Ecole Supérieure du Commerce Extérieur, 10 rue Sextius Michel, Paris, France

ARTICLE INFO

Article history:
Received 25 February 2013
Received in revised form 27 January 2014
Accepted 9 March 2014
Available online xxxx

Keywords:
Creativity judgment
Training
Learning mechanism

ABSTRACT

Evaluating individual creativity is an important challenge in creativity research. We developed a training module for non-expert judges in which participants learned the definitions of components of creativity and received expert feedback in an interactive creativity judgment exercise. We aimed to test whether and how the training module would increase the reliability and validity of non-expert ratings. Study 1 (N = 79) showed that the training had a positive effect on the test–retest reliability and validity of creativity ratings. Study 2 (N = 126) replicated the results on test–retest reliability and validity but with low absolute values, indicating that trained participants cannot substitute experts. In addition, Study 2 showed that the effect of the training module on the validity of creativity ratings was mediated by increased validity of ratings of novelty and elaboration. The results are discussed in terms of theoretical and practical relevance.

1. Introduction

Despite a growing interest in creativity research, the field is often challenged because of the difficulty in finding reliable and valid methods to assess individual creativity. In the literature various measures of individual creativity, including divergent thinking tests, attitude inventories, ratings by peers, and judgments of products, are described (Lubart, 1994). One of the most used techniques to evaluate individual creativity is the Consensual Assessment Technique (Amabile, 1982, 1996; Hennessey, 1994; Kaufman, Baer, & Cole, 2009; Kaufman, Baer, Cole, & Sexton, 2008; Kaufman, Gentile, & Baer, 2005; Runco, 1989). This technique consists of assessing an individual's level of creativity by assessing the creativity level of certain products made by this individual. The technique is based on giving minimal information about creativity to a group of judges who have to rate the creativity of a set products relative to other products in the sample (Amabile, 1996; Dollinger & Shafran, 2005). Because creativity is a relative construct it is commonly expected that judges should be familiar with the domain, i.e. experts or, at least, gifted novices (Amabile, 1996; Dollinger & Shafran, 2005; Kaufman et al., 2009). Several studies show that experts have a good interrater agreement (Kaufman et al., 2009; Kaufman et al., 2008; Kaufman, Baer, Cropley, Reiter-Palmon, & Sinnett, 2013; Kaufman et al., 2005). A potential difficulty, however, in using this technique is finding and asking experts to evaluate sometimes hundreds of products; this may pose a real challenge to creativity researchers (Cropley & Kaufman, 2012). The aim of the present paper is twofold. From a practical viewpoint we aim to investigate whether it is possible to teach novices how to evaluate the creativity of products made by individuals. From a more conceptual/theoretical viewpoint a parallel aim that we have is to shed more light on the underlying processes of learning to judge creativity. By doing this we aim to understand how exactly people arrive at creativity judgments.

1.1. Previous research

Recently, Dollinger and Shafran (2005) suggested that non-expert judges could become more familiar with a domain of products by exposing them their prototypes. They conducted an experiment in which they compared expert judges' creativity ratings of drawings (made by university students) with trained non-expert judges' ratings of the same drawings. The training module consisted of presenting 16 representative drawings to non-expert judges before they had to evaluate the target drawings. Their results showed that trained judges' and expert judges' ratings loaded on a single principal component; thus expert and trained non-expert judges' mean creativity ratings were highly correlated. They concluded that showing non-expert judges representative drawings previous to the actual creativity judgment can make their creativity ratings more similar to expert judges' creativity ratings. However, because this study had no control group, it is difficult to assess causality and the extent to which their training module increased the validity of non-expert judges' creativity ratings compared to a baseline. Considering the number of participants in their study – 5 expert judges and 5 non-

* Corresponding author at: Institut de Psychologie, 71 Avenue Edouard Vaillant, Boulogne-Billancourt, France.
E-mail address: martinstorme@ymail.com (M. Storme).

http://dx.doi.org/10.1016/j.lindif.2014.03.002
1041-6080/

expert judges – it is even more critical to have a baseline to which they could have compared the trained group. Dollinger and Shafran (2005) suggested that further research should be conducted to generalize their results regarding the possibility to enhance the expertise of non-expert judges with a training module.

According to most creativity researchers (Besemer & O'Quin, 1999; Caroff & Besancon, 2008; Runco & Charles, 1993; Storme & Lubart, 2012) creativity ratings of a product typically involve rating subcomponents of creativity, such as novelty, resolution and elaboration. These subcomponents are used in various scales such as the Creative Product Semantic Scale (CPSS; Besemer & O'Quin, 1999; White & Smith, 2001; White, Shen, & Smith, 2002; O'Quin & Besemer, 2006) and the Creative Solution Diagnosis Scale (CSDS; Cropley & Kaufman, 2012), to judge creative ideas, products or designs. A study conducted by Cropley and Kaufman (2012) suggests that novices can be reliable judges when they are provided with precise criteria to assess creativity. The study aimed to develop the Creative Solution Diagnosis Scale, based on four subcomponents of creativity (relevance, novelty, elegance and genesis). The results showed that novices had a high level of interrater agreement (Cronbach's α ranging between .87 and .98) when rating the same products. Note that in this study the researchers did not assess the validity of the novices' creativity ratings by comparing their ratings with ratings of experts, but nonetheless the study shows that relevance, novelty, elegance and genesis are meaningfully related to each other. It appeared that these four components of creativity loaded on a single factor which the authors called 'functional creativity'.

In sum, previous research suggests that novices may learn to judge creativity as long as they have knowledge of the existing products in the domain to be rated (Dollinger & Shafran, 2005) and knowledge of the subcomponents of creativity (Cropley & Kaufman, 2012). Yet to our knowledge there are no existing studies that directly investigated the role of these subcomponents when novices are taught to judge creativity. In the present paper, we report on two studies in which we implement a new training module that we compare to a control group of individuals who are not trained. To give our trainees an analytical understanding of what creativity is, we provide them with a precise definition of creativity involving the subcomponents of creativity. This allows us to investigate the underlying mechanisms involved in the learning of rating creativity. In addition, our trainees are not only passively exposed to representative prototypes of the products to be judged, as in Dollinger and Shafran's (2005) study; but they also receive expert feedback in an interactive creativity judgment exercise.

Our aim was twofold. First, we investigated whether our training module, compared to a control module, would increase the reliability and validity of creativity judgments. Second, to show the underlying mechanism of learning to judge creativity, we investigated whether the effect of the training could be (partially) explained by the extent that trainees made correct use of the subcomponents of creativity.

1.2. Principles of the training module

In this section we describe the basic principles behind our training module. We will elaborate further on the procedural details of the training module in the Method section. In order to study the effect of training non-expert judges, we developed a specific training module based on children's drawings. The training module consists of two stages. The first stage consists in providing participants with a definition of creativity based on explaining to them that products (in this case drawings) differ in creativity because they differ in elaboration, novelty and resolution. More specifically, they are explained that they should take these three subcomponents into account when deciding on the creativity level of a given drawing. In this first stage trainees are also shown prototypical drawings to familiarize them with different levels of creativity. Thus, the first stage aims to give participants an analytical understanding of creativity and familiarize them with representative drawings of varying levels of creativity.

In the second and last stage participants get an opportunity to exercise judging creativity levels of a new set of drawings. First, they are asked to rate the creativity level of a set of drawings. Then they receive feedback on how well they did. Compared with the first stage, the second stage is a more interactive feedback stage in which participants learn to rate creativity.

To sum up, the training consisted of first presenting participants with a precise definition of creativity, then showing participants prototypical drawings, and finally providing participants with feedback on the accuracy of their creativity judgments in a creativity rating exercise. We expected that the training, compared to a control condition, would improve the reliability and validity of creativity judgments in a final set of drawings, and that this would be mediated by improvements in the validity of judgments of the subcomponents of creativity.

1.3. Overview of the studies

This article reports on the results of two studies which aimed at testing the effectiveness of a training module for judging creativity, intended for non-expert individuals, and to investigate learning mechanisms of creativity judgments. Our first study aimed at investigating whether it is possible to learn to judge creativity, and the size of the effectiveness of our training module. We hypothesized that trained participants would produce more reliable and more valid creativity ratings than non-trained participants. To this end, we assessed whether trained participants agreed more with each other (i.e. had higher interrater reliability) than non-trained participants. We also assessed whether trained participants were more stable over time (i.e. had higher temporal stability) in their creativity judgments than non-trained participants. Finally, we assessed whether the ratings of trained participants, compared to the ratings of non-trained participants, were more reliable and more in agreement with the ratings of expert judges (i.e. had higher validity).

The second study aimed at investigating the underlying psychological mechanisms of learning to judge creativity. Because the training module provided our participants with a precise definition of creativity based on novelty, elaboration and resolution, we expected that the training module would also improve the trained participants' validity of novelty and elaboration judgments. More specifically, we expected that the more valid the judgments of novelty and elaboration are (i.e. the more in agreement with expert judges' evaluations of novelty and elaboration), the more valid the judgments of creativity would be as well. In other words, we hypothesized that increased validity of novelty and elaboration judgments would mediate the relationship between the training module and the validity of creativity judgments.[1]

2. Study 1: global effect of the training on reliability and validity

Participants were randomly assigned to a training or control module. The control module was comparable to the training module to the extent that participants were exposed to the same drawings as in the training module, but participants received no definition of creativity and no feedback on their accuracy in the exercise session. Including this condition to our study design allowed us to rule out alternative explanations for the effect of the training, and assess the extent to which the training had an effect compared to a baseline of non-trained participants.

[1] We decided to limit the number of dimensions to be evaluated by the participants to make the judgment process not too complex. Since the products to be evaluated were children's drawings, we dropped judgments of resolution/usefulness from the ratings.

After the training (or the control) module, participants had to rate the creativity of a set of children's drawings. We hypothesized that,

1) the training module, compared to the control module, would increase inter-individual reliability – i.e. the interrater agreement – of the creativity ratings,
2) the training module, compared to the control module, would increase the intra-individual reliability – the temporal stability – of the creativity ratings,
3) the training module, compared to the control module, would increase the validity – i.e. the agreement with expert ratings – of creativity ratings.

2.1. Method

2.1.1. Participants

All participants were 2nd year college students (66 females, 13 males) participating for course credit points. The mean age of the participants was 20.50 ($SD = 3.21$). Students in psychology were chosen because they are one of the targets of this training module; i.e. if the training is effective, they could be hired and trained in order to evaluate the creativity level of products. Participants first received the training ($N = 39$) or the control module ($N = 40$), and then were asked to evaluate the creativity level of twenty new drawings.

2.1.2. Materials

For the training we used children's drawings. These drawings were made on a computer version of EPoC (Evaluation of the Potential of Creativity; Lubart, Besançon, & Barbot, 2011) during a graphic integrative thinking task. Originally, this task was designed to assess the creativity level of children aged from 6 to 12 years old. Children had to produce original and elaborate drawings with a specific constraint which consisted in the inclusion of four abstract shapes. Based on previous data collection (Lubart et al., 2011), we selected 28 drawings to be used in the training module, and used the same 28 drawings in the control module as well. These drawings were previously indicated by expert judges to be representative of differing levels of children's creativity (Lubart et al., 2011).

2.1.2.1. The training module. The training module consisted of two stages. In the first stage, the familiarization stage, participants were explained that the creativity level of a drawing is characterized by its degrees of novelty, resolution and elaboration. Participants were provided with definitions of novelty, resolution and elaboration based on the items of the CPSS (Besemer & O'Quin, 1999). Specifically, they were explained that a drawing is novel when it is original, surprising and germinal; has high resolution when it is valuable, logical and useful; and elaborate when it is organic, elegant, complex, understandable and well-crafted. Then participants were shown fourteen drawings all at once and were explained that these drawings are representative examples of different creativity levels. Knowing the definition of creativity, participants were asked to watch the drawings to get an idea of the range of creativity among the drawings. All the drawings were presented on the same screen, ranked from the least creative drawing to the most creative one. For each drawing, participants could see how they had been previously graded by experts, on a scale between 1 (not creative at all) and 7 (very creative). Participants could watch the drawings as long as they wanted to. In the second stage, the exercise stage, participants had to take a test with fourteen new drawings. Participants were instructed to evaluate those new drawings' creativity level considering their novelty, resolution and elaboration. All the drawings were presented on the same screen to allow for comparison, and participants rated each drawings' creativity level on a Likert scale ranging from 1 'not creative at all' to 7 'very creative'. After participants completed rating of all the drawings, the creativity ratings given by EPoC's experts appeared alongside each drawing. Participants could thus see the actual creativity level of each drawing.

2.1.2.2. The control module. The control module also consisted of two stages. In the first stage, participants saw exactly the same fourteen drawings as in the first stage of the training condition, but received different instructions. The drawings were presented as examples of the material they would have to rate during the study and no mention of creativity was made at this stage. Importantly, in contrast with participants in the training module, participants in the control module also did not receive a definition of what constitutes creativity: They were just passively exposed to the same fourteen drawings used in the familiarization stage of the training condition. Then in the second stage, participants were exposed to the same set of fourteen drawings that were used in the exercise stage of the training condition. The difference with the training condition was that instead of rating the creativity level of the drawings, participants were now instructed to find the number of differences between a given drawing and a slightly modified version of it. The modified version of the drawing consisted of the same drawing with added shapes by the experimenters. Participants had to indicate the number of differences and received the real number of differences as feedback. All the drawings were presented on the same screen: participants could see for each drawing their own estimation of the number of differences (ranging between 1 and 7) and the real number of differences (ranging between 1 and 7).

In sum, the control module was similar to the training module on two aspects: 1) Participants saw the same twenty-eight drawings; and 2) Participants were involved in an exercise task with feedback. The main difference was that in the control module, participants received no definition of creativity, and the exercise that they did was not intended to teach them to rate creativity. Because in both conditions participants saw the same drawings we could rule out the effect of mere exposure to the drawings on the actual creativity ratings. With these two conditions, we could test the effect of providing precise definitions of creativity and giving expert feedback on the reliability and the validity of creativity ratings.

2.1.3. Procedure

The whole procedure was computerized. The total duration of both modules was between 20 and 30 min. Upon arrival at the lab participants were randomly assigned to the training or the control module. Participants were explained that the experiment was about drawings made by children. No reference to creativity was made. Participants were either alone or with another participant in the cubicle, back to back, so they could not see what the other participant was doing. The instructions were computerized and participants had to remain silent.

2.1.3.1. Assessing participants' creativity judgments. After the training (or control module) the computer program automatically switched to the measurement of the dependent variables. To this end, all participants had to rate the creativity level of twenty new drawings (also selected from the EPoC). Participants from the training condition were instructed to evaluate the twenty drawings in relation to one another using what they learned before during their training. More specifically, they assessed each drawing's creativity on a Likert scale ranging from 1 'not at all creative' to 7 'very creative'. Participants from the control condition were also asked to evaluate the creativity level of the twenty drawings on the same 7-point Likert scale. In both conditions, all the drawings were presented in random order on the same screen to allow for comparison. Finally, to be able to measure the temporal stability of the creativity ratings all participants had to come back four weeks later to rate the creativity of the same drawings again. Participants were directly asked to rate the same twenty drawings, without being trained again.

2.1.3.2. Expert judges. Experts were six art teachers (four women and two men) at the elementary school level. Their average age was 45.33 years old ($SD = 10.05$). All experts had more than ten years of experience working with children. They all studied arts and had at least

participated in one art exhibition themselves. Our expert judges did not take part in our training module. They were asked to rate the creativity of the same twenty drawings in relation to one another on a Likert scale ranging from 1 'not at all creative' to 7 'very creative'. All the drawings were shown on the same screen in random order.

2.2. Results

2.2.1. Interindividual reliability of creativity judgments: interrater agreement

We calculated Cronbach's α of the ratings across the twenty drawings, separately for the trained and non-trained judges, and for the expert judges' creativity ratings. Because Cronbach's α is sensitive to the number of judges (Kaufman et al., 2009; Kaufman et al., 2013), and because the number of expert judges was nearly six times smaller than the number of the trained and non-trained judges, we used a re-sampling technique when computing Cronbach's α for our novice judges (Kaufman et al., 2013). Indeed, interrater reliability increases systematically with increasing numbers of judges (Schmitt, 1996). This jeopardizes the comparison of groups with different sample sizes. To deal with this problem, Kaufman et al. (2013) suggested to calculate Cronbach's α of the bigger group (in the case our group of novice judges) by randomly re-sampling smaller sets of judges from the original sample that are equal to the number of judges in the smaller group (in this case our expert raters). We repeated the re-sampling procedure 10,000 times, and computed Bootstrapped Confidence Intervals (BCI).

The results of this analysis showed that non-trained ($\alpha = 0.78$; 95% BCI = [0.57; 0.90]) and trained judges ($\alpha = 0.91$; 95% BCI = [0.87; 0.95]) did not differ significantly from each other regarding their interrater agreement. Interestingly, both our novice groups also did not differ from our experts ($\alpha = 0.89$; 95% BCI = [0.81; 0.94]), whose level of interrater agreement was nevertheless comparable with reports in previous research on expert judges (Kaufman et al., 2009). In sum, the results suggest that there is neither a significant difference between trained and non-trained novice judges, nor between novices and expert judges regarding the level of interrater agreement.

2.2.2. Intraindividual reliability of creativity judgments: temporal stability

Temporal stability of creativity judgments was computed by correlating for each participant the ratings at the first measurement with the ratings at the second measurement. Because distributions of correlation coefficients are not symmetrical, we normalized the correlation coefficients using Fisher's transformation (Fisher, 1921), which transforms correlation coefficients into Z-values. An independent samples *T*-test revealed that trained participants had significantly higher transformed correlations ($M = 1.09$; $SD = 0.35$) than non-trained participants ($M = 0.70$; $SD = 0.31$), $t(77) = 5.29$; $p < .001$; Cohen's $d = 1.18$. This means that trained participants had significantly higher correlations between the first and second measurements, than non-trained participants In other words, their creativity judgments proved to be more stable over time.

2.2.3. Validity of creativity ratings: agreement with experts

Before proceeding with our analyses, for each drawing, we computed the average experts' rating. Next, we computed for each participant the correlation between their own ratings of each drawing and the average expert rating of the same drawings. Analyses were again conducted with transformed correlation coefficients using Fisher's transformation. An independent samples *T*-test revealed that trained participants' ratings were significantly more valid, than non-trained participants' ratings, $t(65.86) = 4.58$; $p < .001$; Cohen's $d = 1.04$. This means that trained participants rated the twenty drawings significantly more like experts, with average transformed correlations of .75 ($SD = .19$), compared with non-trained participants, who had an average transformed correlation of .49 ($SD = .30$).[2]

2.3. Discussion

As expected, the results revealed that the training module significantly increased the average agreement with expert ratings, and thereby the validity of our trained participants' creativity judgment. Regarding the reliability of our novices' creativity judgments, the training module (compared to the control module) significantly improved the stability of creativity judgments among our novice judges. Yet, the interrater agreement among trained novices did not prove to be higher than the interrated agreement among non-trained novices. Thus, whereas our training significantly increased the validity and intraindividual reliability of our trained novices' creativity ratings, interindividual reliability did not appear to be affected by the training. Our novices even had levels of interrater agreement similar to our expert judges. The latter finding suggests that our novices were from a homogeneous population; indeed they were all students of psychology, with approximately the same age and education, and most were female. We come back to this in the General discussion section.

However, from a theoretical and conceptual perspective, the results of this study show that rating creativity can indeed be learned. Therefore, to better understand how the training module taught our novices to rate creativity, in the next study we investigate the underlying psychological mechanisms involved in learning to rate creativity. Such an understanding of the mechanisms underlying the learning of creativity judgments could not only suggest options to improve future training modules, but also shed light on the nature of creativity judgments, which is of high theoretical and conceptual relevance (Caroff & Besancon, 2008).

3. Study 2: explaining the effect of the training

To further understand how the training module increased the validity of creativity ratings, we conducted a second study. Recall that the training module provides definitions of novelty, resolution and elaboration (the subcomponents of creativity). However, in the previous study we did not assess the novices' ratings of novelty, resolution and elaboration, because we first wanted to establish that the training directly affects the reliability and validity of creativity judgments. Now that we have established this, in the current study we also assess judgments of these subcomponents of creativity, *prior* to actually assessing judgments of creativity. Because creativity judgments depend partly on the judgment of these subcomponents (Caroff & Besancon, 2008; Runco & Charles, 1993; Storme & Lubart, 2012), we expect that the more accurate (i.e. more valid in terms of agreement with expert judges' ratings) those judgments are, the more accurate general creativity judgments should be as well. More specifically, we hypothesized that the positive effect of the training module on the validity of creativity ratings, would be mediated by the increase in validity of the subcomponents of creativity ratings due to the training.

3.1. Method

3.1.1. Participants

All participants were 2nd year college students (115 females, 11 males) participating for course credit points. The mean age of the respondents was 21.33 ($SD = 4.99$). Participants first received the training ($N = 63$) or the control module ($N = 63$), and then were asked to

[2] We conducted this analysis on the data from the second measurement (4 weeks later) as well, in order to investigate the long-term effect of the training module on the validity of creativity ratings. Four weeks later, trained judges still agreed significantly more with experts, than non-trained judges did, $t(77) = 5.17$; $p < .001$.

evaluate the novelty, the elaboration and the creativity level of thirty-one new drawings.

3.1.2. Materials

The training module and the control module were identical to the first study. We used the same 20 drawings as in the first study to measure the dependent and mediator variables, and added 11 other drawings to increase the generalizability of our results. The whole procedure was mostly similar to the procedure of the first study except for some aspects that we detail in the present section.

3.1.3. Procedure

The total duration of both conditions was between 30 and 45 min. Each participant was randomly assigned to the training condition or to the control condition. After receiving the training module (or the control module), participants rated the 31 drawings they had never seen before on three dimensions: novelty, elaboration and creativity.[3] Participants gave the three ratings separately on three consecutive screens. Thus, participants first rated each drawing's novelty and elaboration (the order of which was counterbalanced across participants), and then rated each drawings' creativity, relative to the other drawings on 7-point Likert scales ranging from 1 'not at all (novel, elaborate or creative)' to 7 'very (novel, elaborate or creative)'. We chose to measure the subcomponents of creativity prior to measuring creativity judgments, because we assume that creativity judgments are the result of novelty and elaboration judgments. As in Study 1, all the drawings were presented in random order on the computer screen. After rating each drawing's novelty (or elaboration) participants pressed 'Validate', and the program switched to the next screen. On the second screen the same drawings were again presented, but in another randomized order as on the previous screen. Participants now rated each drawing's elaboration (or novelty) level. On the third and final screen (again with the drawings in a new randomized order), participants rated the creativity level of each drawing. In both conditions, and for each dimension under evaluation, participants were specifically instructed to rate the drawings in relation to one another. Participants from the training condition were explicitly instructed to use what they had previously learned during the training module regarding the definition of creativity and its subcomponents. Participants in the control condition were not asked to do this. Finally, similar to the previous study, we again measured the temporal stability of the creativity ratings; all participants had to come back seven weeks later to rate the novelty, elaboration and creativity of the same 31 drawings again. Participants were directly asked to rate the drawings, without being trained again.

3.1.3.1. Experts. The experts were the same as in the first study, and they rated the same drawings as the participants. The drawings were again presented on the same screen and in random order. Similar to the novices, experts were instructed to rate the elaboration, novelty and creativity of each drawing in relation to one another (also on three consecutive screens) of the same 31 drawings (the order of which was counterbalanced between experts). Experts always rated the creativity level of the drawings first to enable comparability of the results with the results in the previous study.

3.2. Results

3.2.1. Interindividual reliability of creativity judgments: interrater agreement

We used the same re-sampling procedure as in Study 1. The results of this analysis showed that non-trained judges ($\alpha = 0.91$; 95% BCI = [0.84; 0.95]) did not differ from trained ones ($\alpha = 0.93$; 95% BCI = [0.90; 0.95]), and together they did not differ from experts either ($\alpha = 0.89$; 95% BCI = [0.81; 0.94]). The current study suggests that the training module had no observable effect on the level of interrater agreement.[4]

3.2.2. Intraindividual reliability of creativity judgments: temporal stability

An independent samples *T*-test with the transformed correlation coefficients as the dependent variable revealed that trained novices had a significantly higher transformed correlation between the first and second measurements ($M = 0.15$; $SD = 0.11$), and thus were significantly more stable over time, than non-trained novices ($M = 0.06$; $SD = 0.13$), $t(124) = 3.76$; $p < .001$; Cohen's $d = 0.75$.[5]

3.2.3. Validity of the creativity ratings

Results regarding validity were also replicated. As in Study 1, we computed the correlation between each participant's rating and average experts' ratings of the same drawings, and coefficients were normalized. An independent samples *T*-test revealed that trained participants' ratings were significantly more valid, than non-trained participants' ratings, $t(124) = 5.77$; $p < .001$; Cohen's $d = 1.01$. This means that trained participants rated the 31 drawings significantly more in agreement with experts, with average transformed correlations of .22 ($SD = .11$), compared with non-trained participants, who had an average transformed correlation of .11 ($SD = .11$).[6]

3.2.4. The mediating role of novelty and elaboration ratings

Finally, we tested the two hypothesized mediators which could explain how the training module increased the validity of creativity ratings. It was hypothesized that the training module exerted influence on the validity of creativity ratings through increasing the validity of novelty and elaboration ratings. To conduct this two-mediator analysis, we followed the recommendations of Preacher and Hayes (2008). When one has more than one mediator variable, it is preferable to use a single integrative analysis than two simple mediation analyses. We used Preacher and Hayes (2008) SPSS macro, in order to compute the direct and indirect effects. Their method allows the computation of a total indirect effect, along with indirect effects of each of the mediators controlling for the other ones. In a multi-mediator model the total indirect effect is the sum of the specific indirect effects of each suggested mediator.

We entered the mean correlation coefficients of creativity (i.e. validity of creativity ratings) as the dependent variable, the training module (dummy coded: 1 = training module; 0 = control module) as the independent variable, and the mean correlation coefficients of novelty and elaboration ratings (i.e. validity of novelty and elaboration ratings) as the mediators. The mean correlation coefficient of novelty and elaboration ratings was computed the same way as the mean correlation coefficients of creativity ratings. For both novelty and elaboration, we computed for each participant the correlation between their own rating and the average expert rating. Correlation coefficients were then normalized using Fisher's transformation.

For the mediation analysis we followed Preacher and Hayes' (2008) recommendation to use bootstrapping. For this bootstrap procedure, we set the number of samples to 1000. The model and the estimates of the

[3] Again we did not take into account resolution in order to limit the number of dimensions under investigation.

[4] Interrater agreement for ratings on novelty and elaboration was comparable to those obtained on creativity ratings. Experts showed satisfactory interrater agreement on both ratings (novelty: $\alpha = 0.92$; 95% CI = [0.87; 0.96], elaboration: $\alpha = 0.93$; 95% CI = [0.89; 0.97]), and trained novices did not differ significantly (novelty: $\alpha = 0.89$; 95% BCI = [0.84; 0.93], elaboration: $\alpha = 0.92$; 95% BCI = [0.88; 0.95]) from non-trained novices (novelty: $\alpha = 0.88$; 95% BCI = [0.71; 0.93], elaboration: $\alpha = 0.91$; 95% BCI = [0.86; 0.95]), nor from experts.

[5] Trained participants, compared to non-trained participants, were also more stable over time regarding elaboration ratings, $t(124) = 3.88$; $p < .001$. The difference in novelty ratings between trained and non-trained participants did not reach significance, although the means were in the expected direction, $t(124) = 1.37$; $p = .17$.

[6] Seven weeks later, the difference between trained and non-trained judges regarding the validity of creativity judgments had disappeared, $t(124) = 0.70$; $p = .48$.

direct and indirect effects are reported in Fig. 1 and the detailed statistics are reported in Table 1.

The analysis indicated that the training module significantly increased the validity of novelty ($B = .05$; $p < .05$) and elaboration ratings ($B = .08$; $p < .01$). The validity of creativity was significantly increased by the validity of novelty ($B = .20$; $p < .05$) and elaboration ratings ($B = .30$; $p < .01$) as well. Furthermore, this analysis revealed that the total indirect effect was significantly different from zero ($B = .03$; $p < .01$), suggesting that the effect of the training on the validity of creativity was mediated by novelty and elaboration. The separate indirect effects of novelty ($B = .01$; $p < .05$) and elaboration ratings ($B = .02$; $p < .01$), showed that both mediators contributed significantly to the total mediation effect. The mediation was only partial because the direct effect of the training module on the validity of creativity ratings is still significant when controlling for the two mediators ($B = .08$; $p < .01$).

3.3. Discussion

Our second study replicated the results of Study 1 and provided clues about the mechanism behind the learning process of judgments of creativity. Like in Study 1, trained participants had significantly more valid creativity ratings compared with non-trained participants, and the effect size was comparable as the one found in Study 1. Moreover, the training module, compared to the control module again significantly increased the stability of creativity ratings over time among our novices. Also similar to Study 1, the training module did not increase significantly the level of interrater agreement, and the interrater agreement of novices was again comparable with the interrater agreement among expert judges. Note, that absolute levels of validity and temporal stability were lower in Study 2 compared with Study 1.

The training module allowed us to demonstrate how novices learn to judge creativity. As expected, the training module increased the validity of novelty and elaboration ratings, which in turn increased the validity of creativity ratings. In conclusion, the positive effect of our training module on the validity of creativity ratings is explained by its positive influence on novelty and elaboration ratings.

4. General discussion

In the present paper, we investigated whether a training module, compared to a control module, would increase the validity and reliability

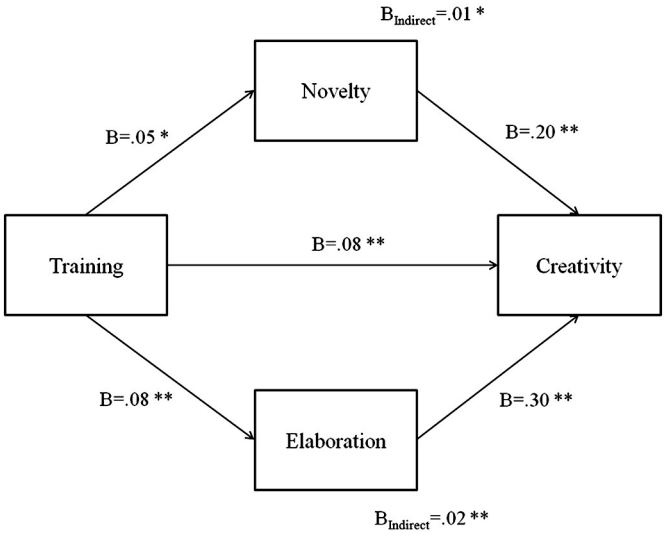

Fig. 1. Effect of the training module on the validity of creativity ratings through the validity of originality and elaboration ratings. *Note.* *$p < .05$; **$p < .01$.

Table 1
Mediation of the training module on the validity of creativity ratings through the validity of originality and elaboration ratings: Unstandardized estimates of the indirect effects are estimated with Bootstrapping (N = 1000).

Effect	Path	Estimate	*SE*	*p*
Direct	Training → Creativity	.08	.02	<.01
	Training → Novelty	.05	.02	.03
	Training → Elaboration	.08	.02	<.01
	Novelty → Creativity	.20	.08	.01
	Elaboration → Creativity	.30	.10	<.01
Indirect	Total	.03	.01	<.01
	Through novelty	.01	.01	.03
	Through elaboration	.02	.01	<.01

of creativity judgments of non-experts. We also studied the underlying mechanism of learning to judge creativity, by investigating whether the effect of the training could be explained by the extent that trainees made correct use of the subcomponents of creativity judgments. These two aims have respectively practical and theoretical relevance.

From a theoretical viewpoint, our studies revealed that it is possible to teach novices how to rate creativity more like experts, as was apparent from significantly higher validity scores among trained novices, compared to non-trained novices. Moreover, the training also increased the intraindividual reliability (i.e. temporal stability) of novices' creativity judgments. The first study not only showed that trained non-experts could rate creativity significantly more like experts, but it also showed that trained judges' ratings were more stable over time compared to non-trained judges' ratings. The second study allowed us to go further in understanding this effect. It appeared that the training module increased the validity of novelty and elaboration ratings, which, in turn, increased the validity of creativity ratings. Similar to Study 1, the training module again appeared to increase the temporal stability of the creativity judgments compared to the control module.

The above findings are relevant from a theoretical perspective, because they show us that rating creativity can be learned and how it can be learned. Such an understanding of learning mechanisms is important when we want to understand the nature of creativity judgments. In existing literature, people who are considered skilled in rating creativity are either experts (Kaufman et al., 2009; Kaufman et al., 2008; Kaufman et al., 2013) or people who are creative themselves (Caroff & Besancon, 2008). It is assumed that these individuals have a better understanding of what composes creativity. The findings of the present paper suggest that creativity judgments are composed of judgments of novelty and elaboration, and that lay people can be taught what creativity is by teaching them that they should pay attention to these criteria when judging products for their creativity.

From a practical viewpoint however, the absolute value of agreement between our trained novices with our experts might be considered too low to present the training module as an alternative to expert ratings. This is especially the case considering the results of Study 2 in which the average correlation between trained novices and experts did not exceed .22. Yet in Study 1, the average correlation between trained novices and overall expert ratings was considerably higher (.75). One could contend that this finding is related to the order in which the different dimensions were rated. Experts rated creativity first, and then the other two dimensions, whereas this was the other way around for the novices. This is because we consulted experts once, *after* both studies had been conducted. We aimed at getting the experts' initial intuitive judgment of creativity, and therefore asked them to rate creativity first. Because of this, in Study 2, our novices possibly engaged in a different cognitive process than our experts. In other words, if our experts had rated novelty and elaboration first, just like the novices, the validity levels could have been more similar to the ones that we found in Study 1. However, considering the fact that our novices in Study 2 were less stable over time in their creativity judgments, compared to our novices in Study 1, it seems that in Study 2 our novices

somehow had more difficulty in rating creativity. Therefore, we deem it more likely that the relatively more analytical cognitive investment involved in rating novelty and elaboration (prior to rating creativity) went at the cost of the relatively more intuitive process of judging creativity.

Note that the positive aspect of Study 2 is that it gives us insights in the cognitive process of judging creativity – as novelty and elaboration ratings significantly predicted creativity judgments – but the results also suggest that when people judge creativity they should do so in an intuitive manner. In this sense Study 2 is not an exact replica of Study 1, and one should interpret the differences regarding validity with caution. The lower validity scores in Study 2 could be either due to the fact that it is not possible to turn novices into experts, or the order in which we assessed our dependent variables.

In sum, although our training module proved to be effective (compared to our control module), at this point we cannot conclude that we can easily replace experts with trained novices. Nevertheless, considering the relatively short and inexpensive nature of our training module, future endeavors could investigate different and possibly more effective teaching strategies. Perhaps a longer training spread over several sessions, could improve the effectiveness. The present studies imply that one should take extra caution in deciding on the exact dependent variables depending on the research aim. Especially when the aim is practical in nature – i.e. to test whether a training module works – we recommend to only assess creativity judgments as the prime dependent variable, as these judgments might be especially unstable among novices and easily affected when other dimensions are rated as well.

An unexpected finding was that our studies provided no evidence that the training module increased interrater agreement among our novices. Moreover, the absolute level of interrater agreement was very high and comparable to the level of interrater agreement among our experts (>.90). Our novices were from a homogeneous population which could explain the extent to which they agreed with each other. Thus it seems that our training was not strong enough to increase an already existing high level of consensus among our novices. Nonetheless, our training did improve the stability of creativity judgments *within* individuals. Thus although the training did not improve inter-individual reliability, it did improve intra-individual reliability.

Note that, whereas interrater agreement is very important for experts, because it is a check on their agreement in the specific domain to be rated, for non-experts a high interrater agreement alone cannot guarantee any validity of their creativity judgments. Therefore, a validity check with experts is especially crucial when focusing on novices, because it is the only evidence that we can have to demonstrate that novices actually learned to judge creativity like experts. As is evident from their lower agreement with experts, our non-trained judges seemed to agree on something else than the real creativity of the drawings. Thus a high interrater agreement tells us that judges agree with each other, but not on *what* they agree. Therefore, we believe that the emphasis should be more on agreement with experts.

4.1. Strengths, limitations and future directions

A strength of our research lies in the use of an experimental design, involving a comparable control condition. This allowed us to isolate the effect of the training module from other variables. Furthermore, our research did not only focus on interrater agreement but also focus on temporal stability and validity. Previously, Storme and Lubart (2012) showed in a correlational study that novelty, resolution and elaboration judgments were predictors of creativity judgments. Our research adds to their findings and is the first to use a learning paradigm in showing that novelty and elaboration are part of the complex process of judging creativity.

Although our studies provide interesting results, they have limitations as well that could be addressed in further research. Future research could use participants from a more heterogeneous population (in age and gender), as this could increase the generalizability of findings. It would be interesting to replicate the studies on other populations, especially older populations, which could respond differently to the training module. The training could also be adapted for other products made by children, such as stories, or products designed by adults. A stronger test of our idea would be to use products with which lay people are less familiar with. One could contend that most people are familiar with children's drawings and can therefore easily improve their creativity judgments with a short training. Training for example Western lay people to judge the creativity of traditional indigenous masks from Oceania might be much more difficult.

Finally, our studies did not take into account resolution as a predictor of creativity judgments, which is also an important feature of creativity. Providing that the training module gives a definition of resolution, it could be hypothesized that resolution judgments also mediate the relationship between the training module and the validity of creativity ratings.

As a conclusion, the results of the present paper show that in principle it is possible to teach lay people to judge creativity more like experts, at least in the area of children's drawings, and suggest ways to design a training module based on the subcomponents of creativity judgments. However, our results also call for more research to investigate the boundary conditions and generalizability of this effect, and perhaps more importantly whether it is possible to turn novices into experts quicker than the time that we need to find experts.

References

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, *43*(5), 997–1013.

Amabile, T. M. (1996). *Creativity in context: Update to 'the social psychology of creativity'*. Westview Press.

Besemer, S. P., & O'Quin, K. (1999). Confirming the three-factor creative product analysis matrix model in an American sample. *Creativity Research Journal*, *12*(4), 287–296.

Caroff, X., & Besancon, M. (2008). Variability of creativity judgments. *Learning and Individual Differences*, *18*(4), 367–371.

Cropley, D. H., & Kaufman, J. C. (2012). Measuring functional creativity: Non-expert raters and the Creative Solution Diagnosis Scale. *The Journal of Creative Behavior*, *46*(2), 119–137.

Dollinger, S. J., & Shafran, M. (2005). Note on consensual assessment technique in creativity research. *Perceptual and Motor Skills*, *100*, 592–598.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32.

Hennessey, B. A. (1994). The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal*, *7*, 193–208.

Kaufman, J. C., Baer, J., & Cole, J. C. (2009). Expertise, domains, and the consensual assessment technique. *Journal of Creative Behavior*, *43*(4), 223–233.

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and non-expert raters using the consensual assessment technique. *Creativity Research Journal*, *20*(2), 171–178.

Kaufman, J. C., Baer, J., Cropley, D. H., Reiter-Palmon, R., & Sinnett, S. (2013). Furious activity vs. understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 332–340, http://dx.doi.org/10.1037/a0034809.

Kaufman, J. C., Gentile, C. A., & Baer, J. (2005). Do gifted student writers and creative writing experts rate creativity the same way? *Gifted Child Quarterly*, *49*(3), 260–270.

Lubart, T. (1994). Creativity. In R. J. Sternberg (Ed.), *Thinking and problem solving. Handbook of perception and cognition* (pp. 289–332). Academic Press.

Lubart, T., Besançon, M., & Barbot, B. (2011). *Evaluation du potentiel créatif (EPoC)*. Hogrefe.

O'Quin, K., & Besemer, S. P. (2006). Using the creative product semantic scale as a metric for results-oriented business. *Creativity and Innovation Management*, *15*(1), 31–41.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891.

Runco, M. A. (1989). The creativity of children's art. *Child Study Journal*, *19*, 177–190.

Runco, M. A., & Charles, R. E. (1993). Judgments of originality and appropriateness as predictors of creativity. *Personality and Individual Differences*, *15*(5), 537–546.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353.

Storme, M., & Lubart, T. (2012). Conceptions of creativity and relations with judges' intelligence and personality. *The Journal of Creative Behavior*, *46*(2), 138–149.

White, A., Shen, F., & Smith, B. L. (2002). Judging advertising creativity using the creative product semantic scale. *Journal of Creative Behavior*, *36*(4), 241–253.

White, A., & Smith, B. L. (2001). Assessing advertising creativity using the creative product semantic scale. *Journal of Advertising Research*, *41*(6), 27–34.