The relationship between methods of scoring the alternate uses task and the neural correlates of divergent thinking: Evidence from voxel-based morphometry

Oshin Vartanian\textsuperscript{a,b,}\textsuperscript{*}, Ingrid Smith\textsuperscript{b}, Timothy K. Lam\textsuperscript{a}, Kristen King\textsuperscript{a}, Quan Lam\textsuperscript{a}, Erin L. Beatty\textsuperscript{a,c}

\textsuperscript{a} Defence Research and Development Canada, Toronto Research Centre, 1133 Sheppard Avenue West, Toronto, ON M3K 2C9, Canada
\textsuperscript{b} University of Toronto, Canada
\textsuperscript{c} University of Southern Denmark, Denmark

\textbf{A B S T R A C T}

Divergent thinking tests have been used extensively in neuroscientific studies of creativity. However, output from tests of divergent thinking can be scored in different ways, and those scores can influence assessments of divergent thinking performance and its relationship with brain activation. Here we sought to investigate the relationship between various methods of scoring the Alternate Uses Task (AUT)—a well-known test of divergent thinking—and regional grey matter volume (GMV) using voxel-based morphometry (VBM). We assessed AUT performance based on (a) traditional approaches that involve scoring participants’ output on fluency, flexibility, originality, and elaboration, (b) a subjective approach that involves scoring output directly on “snapshot” creativity, and (c) the definitional approach that involves scoring output separately on novelty and usefulness—the two criteria deemed necessary and jointly sufficient to categorize an idea as creative. Correcting for age, sex, intracranial volume, verbal IQ and working memory capacity, we found negative correlations between regional GMV in the left inferior temporal gyrus (ITG) and novelty and usefulness scores, but no correlation involving other scoring approaches. As part of the brain’s core semantic system, this region is involved in concept retrieval and integration. We discuss the implications of these findings for our understanding of the neural bases of divergent thinking, and how ITG could be related to the generation of novel and useful responses.

\textbf{1. Introduction}

Divergent thinking is defined as the ability to generate multiple solutions in response to a given stimulus or problem (Guilford, 1967), wherein the process taps “cognition that leads in various directions” (Runco, 1999, p. 577). Historically, tests of divergent thinking have been used extensively to study the psychological underpinnings of creativity (Plucker and Renzulli, 1999). In essence, the use of divergent thinking tests to assess creativity rests on the assumption that creativity benefits from the capacity for idea generation. Indeed, divergent thinking has been shown to predict creative achievement (Guilford, 1966; Kim, 2008; Plucker, 1999; Runco et al., 2010). Importantly, however, the ability to generate ideas reflects only one aspect of the multifaceted construct of creativity. Specifically, within Guilford’s (1967) Structure of Intellect (SOI) model, divergent production is one of five operations (i.e., cognition, memory, divergent production, convergent production, and evaluation) that contribute to intelligence. In other words, divergent thinking is only one of many types of cognition that underpins intelligent behaviour, including creative aspects thereof (see also Reiter-Palmon, 2018).

Reflecting deep concerns within the literature regarding the assessment of creativity (see Barbot and Reiter-Palmon, 2019; Plucker and Makel, 2010), an important question that arises in relation to the measurement of divergent thinking ability is how to score its output (Plucker and Runco, 1998). In a recent systematic review of the scoring procedures for tests of divergent thinking, Reiter-Palmon et al. (2019) identified many factors that can affect its assessment, including the choice of dimension (e.g., originality vs. frequency of responses), instruction-scoring fit (i.e., the extent to which task instructions reflect assessment goals), adequacy of responses (e.g., whether to score all responses or to exclude incomplete, nonsensical, or repetitive responses), objectivity (vs. subjectivity), level of scoring (response vs. ideational pool level), and the method of aggregation, among others. Given these factors, it is easy to see that assessments of perfor-
mance can vary greatly even when researchers are using the same test, let alone different measures for divergent thinking.

Perhaps not surprisingly, divergent thinking tasks have also been used extensively in neuropsychological studies of creativity (see Boccia et al., 2015; Gonen-Yaacovi et al., 2013; Wu et al., 2015). A recent review of the literature revealed that over half of the published studies in the neuropsychological literature had used divergent thinking tasks as the outcome measure of interest (Benedek et al., 2019). It is likely that in the context of neuropsychological investigations, the scoring of divergent thinking tests is influenced by the same factors as those in behavioural studies. In addition, there is also the possibility that the neural correlates we observe in relation to divergent thinking might be influenced by our varying approaches to scoring the tests.

Recently Vartanian et al. (2019) addressed this issue by reanalyzing data from Vartanian et al. (2018a) in which participants had been administered the Alternate Uses Task (AUT) in the functional magnetic resonance imaging (fMRI) scanner. The AUT involves asking participants to generate as many novel uses as possible in response to verbal prompts (e.g., brick). Using Fink et al.’s (2009) paradigm, the control condition involved asking participants to recall as many characteristics as possible in response to the same prompts. For example, the physical characteristics that characterize “knife” could be sharp and metallic, amongst others. There was no need for the physical characteristics to be novel. Methodologically, when inside the fMRI scanner the participants entered their responses (i.e., number of generated uses or recalled characteristics depending on the condition) using fMRI-compatible response buttons. Because generating uses and recalling characteristics both necessitate retrieval from long-term memory, but only the former also involves the requirement to generate novel instances, the contrast between the two conditions is used to isolate the neural correlates of novel idea generation. The results of Vartanian et al. (2018a) had revealed that divergent thinking (compared to recalling characteristics) was correlated with greater activation in the left posterior cingulate cortex (PCC). This finding was consistent with recent work demonstrating the critical role played by PCC in divergent thinking, specifically in terms of mediating functional coupling and transitions between different brain networks that underlie divergent thinking (Beatty et al., 2016; Beatty et al., 2018), likely in support of facilitating the phase-dependent contributions of controlled and goal-directed thinking to creative idea generation (see Chrysikou, 2018; Volle, 2018). Shortly after exiting the fMRI scanner the participants were administered a subset of the same items that they had encountered inside the fMRI scanner and instructed to write down their responses, which were in turn scored in relation to various approaches to scoring the AUT (e.g., counting the total number of generated responses, judging the creativity of the overall set of responses, etc.). Although psychometrically we had observed high and statistically significant correlations amongst most scoring approaches to the AUT, regression analyses demonstrated that PCC activation predicted variation in only a small subset of scoring approaches. This suggests that the predictive utility of a neural index (e.g., left PCC activation) for divergent thinking performance as measured by the AUT can vary as a function of the specific scoring procedure under consideration. As such, at least within functional studies, it appears that our conception of the neural correlates of divergent thinking is influenced by the specific ways in which we opt to score test output. The motivation behind the present study was to see whether the method by which the AUT is scored can influence our conception of the structural neural correlates of divergent thinking.

1.1. Methods of scoring the alternate uses task (AUT)

There are numerous ways to score the AUT (see Plucker and Renzulli, 1999; Reiter-Palmon et al., 2019). Here we will focus on three relatively common scoring approaches. First, traditional approaches involve scoring each participant’s output based on indices such as fluency (overall sum of generated uses), originality (statistical infrequency of generated uses), flexibility (number of conceptual categories within which uses could be binned), and elaboration (degree of detail and richness in a response), amongst others (Guilford, 1967). Second, in part motivated by Amabile’s (1982) Consensual Assessment Technique (CAT), researchers have advocated for the use of several subjective scoring approaches according to which a panel of experts rates the generated uses directly on creativity or related constructs such as idea quality (see Silvia et al., 2008). These subjective scoring techniques can take various forms, such as scoring the overall output of a participant on creativity (i.e., “snapshot” approach), or scoring each individual response separately on creativity and then averaging across responses (see Reiter-Palmon et al., 2019). Third, researchers have adopted definitional approaches according to which the output of divergent thinking tasks is scored separately on novelty and usefulness—the two criteria deemed necessary and jointly sufficient to categorize an idea as creative (Diedrich et al., 2015). Importantly, there is reason to suspect that the scores represent different sets of mental processes and operations that underlie performance, reflected by their variable correlations with measures of intelligence and executive functions (Benedek et al., 2012; Benedek et al., 2014; Jauk et al., 2013; 2014; see also Zabelina et al., 2019).

Here, it is important to acknowledge two important issues regarding any measure used to score the AUT. First, in all likelihood, none of the scoring approaches for the AUT reflects process specificity in the strict sense of the word. In other words, it is not possible to isolate a single cognitive process and/or mechanism that underlies individual differences in performance reflected by any score. Consider the contributions of the capacities and abilities associated with fluency versus originality ( Cotter et al., 2019): Fluency has been associated with retrieval ability (Beaty et al., 2014; Silvia et al., 2013), processing capacity and speed (Ruhn and Holling, 2009), and the ability to assess the relatedness between terms quickly (Vartanian et al., 2009). In turn, originality has been associated with problem-finding ability (Abdulla et al., 2020), the ability to generate semantically distant synonyms of cue words (Beaty et al., 2014), and fluid intelligence (Beaty et al., 2014). Therefore, any difference one might find in the functional/structural neural correlates of fluency versus originality cannot be attributed to a single process and/or mechanism exclusively. Having said that, it is nevertheless possible to consider how a constellation of the associated abilities and capacities contributes to each score, and to consider similarities and differences between scoring methods in that regard. In the case of fluency and originality, one can argue that fluency loads relatively more heavily on a constellation of processes that involve speed of retrieval from long-term memory, whereas originality in addition loads relatively more heavily on a constellation of processes that draw on intelligence and goal-directed behaviour (i.e., executive functions).

Second, and related to the first point, the various scoring approaches to the AUT do not exhibit theoretically and/or psychometrically independent associations with the construct of creativity. Rather, they reflect varying but interdependent ways of measuring this multifaceted construct. For example, it is now generally accepted that for an idea to qualify as creative, it must be perceived as both novel and useful. Defined as such, it would appear that novelty and usefulness are two independent dimensions for assessing creativity. However, Diedrich et al. (2015) have shown that in deciding whether an idea is creative or not, participants take usefulness into account only after they have determined that the idea is novel to begin with. In this sense, novelty has precedence in the processing queue for determining creativity. This does not mean that novelty and usefulness cannot be measured in-

---

1 Expanding on this earlier body of work, Snyder and colleagues used principles derived from information theory to compute the creativity quotient—a composite measure that accounts for the number of ideas as well as their flexibility (Snyder et al., 2004).

2 Arguably, one could also include “surprise” as another criterion (see Simonton, 2012).
Table 1

Studies examining the correlation between creativity and grey matter (volume, concentration, density) using VBM in neurologically healthy human samples (Xia et al., 2017; Xurui et al., 2018; Yang et al., 2017; Zhu et al., 2016; Zhu et al., 2013; Bendetowicz et al., 2017; Chen et al., 2014; Fink et al., 2014; Gansler et al., 2011; Arkin et al., 2019; He et al., 2018; Li et al., 2019; Li et al., 2015; B. Shi et al., 2017; B. Shi et al., 2017; Takeuchi et al., 2010).

| First author | Year | n   | Creativity Task                               | Creativity index                                  | Positive                                      | Negative                  |
|--------------|------|-----|-----------------------------------------------|---------------------------------------------------|-----------------------------------------------|---------------------------|
| Takeuchi     | 2010 | 55  | S-A Creativity Test                           | Composite score                                   | Midbrain, caudate, middle frontal gyrus, precuneus, caudate | None                      |
| Gansler      | 2011 | 18  | Figural TTCT                                  | Composite score                                   | Parietal lobe IFG                             | None                      |
| Zhu          | 2013 | 285 | Verbal TTCT                                   | Composite score                                   | Superior frontal gyrus, VMPFC, ITG, precuneus, DLPFC, thalamus | None                      |
| Chen         | 2014 | 366 | CAQ                                          | CAQ score                                         | Cuneus and precuneus (BIS), cuneus and occipital gyrus (AUT originality) | None                      |
| Fink         | 2014 | 76  | Verbal imagination subscales of BIS, AUT      | Composite BIS score; fluency and originality of AUT using snapshot method | Precuneus and striatum (originality) | None                      |
| Jauk         | 2015 | 135 | AUT, Instances Task                          | CFA-derived fluency and originality scores        | MTG, ITG                                     | Precentral gyrus, middle frontal gyrus, OFC |
| Li           | 2015 | 246 | CAP                                          | Composite score                                   |                                               |                           |
| Zhu          | 2016 | 163 | CBI, CAQ                                     | CBI and CAQ scores                                | PMC (CBI)                                    | None                      |
| Yang         | 2017 | 270 | Verbal TTCT                                   | Composite score                                   | Supramarginal gyrus                         | None                      |
| Shi          | 2017a| 356 | CAQ                                          | CAQ Artistic and Scientific subscales             | Middle frontal gyrus, inferior occipital gyrus (Scientific creativity) | None                      |
| Shi          | 2017b| 356 | Verbal TTCT                                   | Originality score                                 | TOJ (females)                                | TOJ (males)               |
| Bendetowicz  | 2017 | 54  | Remote Combination Association Task           | Performance, solution by insight, and appropriateness |                                           |                           |
| Xia          | 2017 | 64  | CAP                                          | Composite and subscale scores                      | OFC and DLPFC (overall score), OFC (Imagination subscale) |                           |
| Xurui        | 2018 | 80  | Figural TTCT                                  | Composite score                                   | Medial frontal gyrus and ACC (Art majors)     |                           |
| He           | 2018 | 213 | CAP                                          | Composite and subscale scores                      | OFC (composite and Imagination, Curiosity, Challenge, and Risk-taking subscales) |                           |
| Li           | 2019 | 261 | RAT                                          | Accuracy                                        | None                                         | ACC                       |
| Arkin        | 2019 | 38  | ICT                                          | Creativity and/or Imagination                      | Hancockpus, ITG                             |                           |

Notes. VBM = Voxel-based morphometry; n = sample size; AUT = Alternate Uses Task; BIS = Berliner Intelligenz-Struktur-Test; CAP = Williams Creativity Assessment Packet; CBI = Creative behaviour Inventory; CFA = Confirmatory actor analysis; CAQ = Creative Achievement Questionnaire; RAT = Remote Associates Test; TTCT = Torrance Test of Creative Thinking; ACC = Anterior cingulate cortex; ICT = Improvisation Continuation Task; IFG = Inferior frontal gyrus; DLPFC = dorsolateral prefrontal cortex; ITG = Inferior temporal gyrus; MTG = Middle temporal gyrus; OFC = Orbitofrontal cortex; PHG = Parahippocampal gyrus; PMC = Premotor cortex; SMA = Supplementary motor area; RLPFC = Rostrolateral PFC; STG = Superior temporal gyrus; TOJ = Temporo-occipital junction; VMPFC = ventromedial prefrontal cortex; Creativity task = Creativity measure administered to the participants; Creativity index = Actual measure/score used in the VBM analyses.

1.2. Present study and hypothesis

There exists substantial literature examining the correlation between creativity and variations in brain structure involving structural neuroimaging approaches such as structural MRI, diffusion tensor imaging (DTI) and proton magnetic resonance spectroscopy (1H-MRS), as well as lesion studies (for review see Jung et al., 2013; Takeuchi and Kawashima, 2018). Within this larger literature, there has been significant work conducted using voxel-based morphometry (VBM) to measure the correlation between variation in regional grey matter volume (GMV) and various measures of creativity, including divergent thinking, in neurologically healthy samples (Table 1).3 Referencing this literature, we used VBM to measure the correlation between regional GMV and AUT performance scored in seven different ways: (1) fluency, (2) flexibility, (3) originality (4) elaboration, (5) snapshot creativity, (6) novelty, and (7) usefulness. Fluency, flexibility, originality and elaboration were investigated because they represent historically important ways to assess the creativity of responses both in the Torrance Tests of Creative Thinking (TTCT, Torrance, 1966) as well as in Guilford’s (1967) SOI model. In terms of those scoring approaches, fluency, flexibility and originality have been used more frequently than elaboration. However, it was recognized early by Torrance and Guilford that elaboration has an important role to play in creativity by contributing to the development and refinement of ideas. Elaboration was included here because we believe that it plays a complementary role to processes that serve novelty generation, by refining and developing ideas leading to response generation. Snapshot creativity was included as a representative of more recent subjective scoring approaches (Silvia et al., 2008), although we are aware that other approaches to subjective scoring also exist (see Benedek et al., 2013). Finally, novelty and usefulness were included because they represent the two core dimensions of our current scientific definition of creativity (see Kaufman and Sternberg, 2019). We hypothesized that each

3 We acknowledge that there are studies that have investigated the association between regional white matter volume and creativity (e.g., Takeuchi et al., 2017). However, the focus in our study was on regional GMV, in part because the higher frequency of GMV studies allows a better review of variation in scoring approaches to creativity testing.
measure would be correlated positively and/or negatively with regional GMV in different areas of the cortex. We also used a conjunction analysis to see whether any region of the brain might exhibit an invariant pattern of covariation with AUT performance across scoring approaches. In turn, examining the functional contributions of such regions could reveal cognitive processes and/or mechanisms that are commonly recruited to support divergent thinking irrespective of the specific scoring procedure under consideration.

2. Materials and methods

2.1. Participants

The participants were 44 neurologically healthy right-handed volunteers (31 males, 13 females) with normal or corrected-to-normal vision from Canada’s Department of National Defence (NDN) who volunteered to participate in the study. Handedness was assessed using a standard self-report questionnaire (Oldfield, 1971). No participant reported colour blindness. The participants ranged from 20 to 56 years of age ($M = 35.47 ± 11.3$ years). Age was normally distributed, Kolmogorov-Smirnov = 0.12, $p = .11$. From a sampling perspective, the age range of the sample was an unbiased representation of this population where 80% of its members are between the ages of 25 and 54 years (Park, 2007). Regarding gender distribution, the Canadian military is predominantly male, with female representation at 15.90% (https://www.canada.ca/en/department-national-defence/services/women-in-the-forces/statistics.html). As such, at 29%, females were overrepresented in the present sample.

The protocol for the study was approved by the Human Research Ethics Committee of Defence Research and Development Canada, and by the Research Ethics Board of Sunnybrook Health Sciences Centre. The structural MRI scans were collected in the same session but in advance of the functional MRI scans reported in Vartanian et al. (2018a). The analyses reported here are based on 41 participants who completed the entire protocol (i.e., MRI scanning, AUT, intelligence and working memory span measures).

2.2. Materials and procedures

The participants completed one of two 5-item sets of the AUT under standard paper-and-pencil laboratory conditions that involved a 3-minute time limit per object (Vartanian et al., 2007; 2009). Each participant's output was scored for (1) fluency, (2) flexibility, (3) originality (unique = 1, all else = 0), (4) elaboration, (5) snapshot creativity (based on overall output), (6) novelty (scored at the individual response level and subsequently averaged), and (7) usefulness (scored at the individual response level and subsequently averaged). Scoring of flexibility, elaboration, snapshot creativity, novelty, and usefulness was conducted by two independent researchers in our lab. Our two independent raters were 29 and 36 years old at the time of rating (average age = 32.5 years old), which is comparable to the mean age of the sample ($M = 35.47$).

The specific instructions for scoring were as follows:

(i) Fluency. One of the raters counted all the responses generated by each participant into a total sum score across all five prompts (see Vartanian et al., 2007, 2009). For analyses we calculated the average fluency score generated in response to the five prompts for each participant. Note that for scoring fluency we included all uses regardless of whether they were valid uses or not. This is in contrast to instructions provided in the AUT manual according to which “A use, to be acceptable, should be possible for the object” (Guilford et al., 1960, p. 30; see also Abraham, 2016). According to that criterion, an invalid use is not considered further. There are two reasons why we took all uses into consideration, regardless of whether they were possible/valid or not. First, we were interested in deriving a measure of fluency that was more comparable to measures of verbal fluency derived from tests of semantic memory, in which participants are instructed to generate members from a given semantic category as quickly as possible, and the total number of words generated in a certain amount of time is used as a global indicator of performance (see Mayr, 2002). Given that our task was also timed (i.e., three minutes per prompt), this would enable us to derive a measure of verbal production irrespective of appropriateness. Second, and related to the first point, we were keen on taking all responses into consideration because we also rated all responses on appropriateness, which is a measure of how feasible and potentially implementable a solution is (see below). As such, we wished not to constrain, a priori, our response pool to only valid responses.

(ii) Flexibility. The instruction given to the raters was to determine the number of conceptual categories that the generated responses could be binned into (Guilford, 1967). To calibrate across raters what a conceptual category meant in the context of this specific exercise, the first author discussed in detail the concept of a “tool” with the three raters, based on a hypothetical example. In terms of the hierarchy of concepts, the raters were instructed to generate conceptual bins that in their joint opinion were equivalent to the level of “tool.”

Based on previous examples from our lab the two raters were instructed to reach agreement between themselves in terms of the features they would focus on to determine category membership before rating flexibility. Two raters coded all participants independently for flexibility. For analyses we calculated the average number of conceptual categories generated in response to the five prompts for each participant.

(iii) Originality. We used the method of uniqueness scoring (see Wallach and Kogan, 1965). One of the raters assigned a value of “1” to each response given to a prompt that was unique (i.e., not generated by any other participant in this sample in relation to that prompt), and a value of “0” to any response that was generated by at least one other participant in this sample. Our originality score was the sum of unique uses across all prompts. Aside from the fact that there are other methods of scoring originality, there are even different methods of uniqueness scoring. For example, infrequency or uniqueness of responses can be defined based on the probability of its occurrence within a sample (e.g., <5% or 10%), and their numbers added to generate a total uniqueness score for each participant (e.g., Runco, 2008; Torrance, 1974). However, regardless of these differences, the uniqueness approach is distinguished by its focus on statistical infrequency of responses. We opted for Wallach and Kogan’s (1965) approach to scoring originality for two reasons: First, that approach rewards responses that are by definition infrequent because they are not generated by any other participant in the sample. They are truly unique. Second, that method of scoring is maximally distinguished from the method of snapshot creativity and novelty scoring that are explicitly driven by more subjective judgments of unusualness and/or infrequency (see below).

(iv) Snapshot creativity. This involved assigning a single, holistic score to a set of responses (Silvia et al., 2009). Toward that end, the first author instructed the three raters to assign a rating of creativity to the overall set of responses generated for each of the five prompts using a 6-point scale (0 = not at all creative, 5 = extremely creative) based on the standard definition of creativity, according to which responses must qualify as both novel and useful to be deemed creative (see Kaufman and Sternberg, 2019). As such, they were to use that 6-point scale to determine the extent to which the overall set of responses generated for each prompt satisfied the standard definition of creativity. For analyses we calculated the average snapshot creativity score in relation to the five prompts.

(v) Elaboration. This score reflected the degree of detail and richness in a given response, and constitutes one of the original variables proposed by Guilford (1967) and Torrance (1966, 1974-) for assessing divergent production. Using instances from previous studies in the lab, the first author provided examples of responses that exhibited...
various levels of elaboration. The first author instructed the three raters to assign a rating of elaboration using a 6-point scale (0 = not at all elaborate, 5 = extremely elaborate). Elaboration was scored at the individual response level. For analyses we calculated the average elaboration score in relation to the five prompts for each participant.

(vi) Novelty and usefulness. We relied on a modification of the criteria provided by Diedrich et al. (2015) for deriving these two scores. First, we defined as novel those responses that would be deemed uncommon and only given by a few people. Thus, after reading each response generated in relation to a given prompt, we instructed our raters to rate how uncommon that idea appeared to them (0 = extremely common, 100 = extremely uncommon). Note that in contrast to originality, where the scoring is determined by statistical infrequency (i.e., uniqueness scoring) and constrained by the responses of the specific sample under consideration, scoring of novelty involves an assessment by the independent raters in terms of what they themselves considered to be uncommon. In turn, we defined as useful those responses that are feasible and can be potentially implemented to solve the problem (0 = not at all useful, 100 = extremely useful). Novelty and usefulness were scored at the individual response level. For analyses we calculated the average novelty and usefulness scores in relation to the five prompts for each participant. Kolmogorov-Smirnov tests demonstrated that the distribution of scores in each of the seven cases was normal (p > .05). The inter-rater reliability (Kappa) for each of those five scoring measures was as follows: flexibility (K = 0.92), elaboration (K = 0.86), snippet creativity (K = 0.93), novelty (K = 0.94), and usefulness (K = 0.67).

In addition, all participants were administered the Shipley-2 measures for fluid (Abstract reasoning and Block design) and crystallized (Vocabulary) intelligence (Shipley et al., 2009), as well as measures of simple working memory span (verbal, visual, and matrix) (see Harrison et al., 2013; Vartanian et al., 2016). For verbal span, four-letter monosyllabic words were presented one at a time on a monitor. After each block of words, participants were prompted by the software to recall the words they saw in the order they were presented. Blocks ranged from 3 to 9 words. For matrix span, participants were presented with a 4 x 4 matrix where one square (out of 16) appeared in red and the rest in white. At the end of each block of matrices, participants were instructed to recall the locations of the red squares in the order they were presented. Blocks ranged from 3 to 9 matrices. For visual span, participants were presented with one arrow at a time that pointed in one of 8 directions. At the end of each block of arrows, participants were instructed to recall the directions of the arrows in the order they were presented. Blocks ranged from 3 to 9 arrows. The computer program provided a detailed description of each task prior to the start, and the experimenter reviewed the instructions and provided an example in each case to the participants. The order of the three span tasks was randomized.

Finally, all participants were administered the Big Five Aspects Scale (BFAS), a personality test that further develops the big five personality traits test model by adding two ‘aspects’ to each trait (DeYoung et al., 2007). Ten items are used to assess each of the ten aspects. Participants rated their agreement with how well each statement described them using a five-point scale ranging from strongly disagree to strongly agree. Scores for each aspect are computed by taking the mean of the corresponding ten items. Although participants completed the entire BFAS inventory, because of our interest in creativity we focused on the two aspects that constitute openness to experience exclusively: Openness and Intellect. At the aspect level, Intellect reflects perceived intelligence and intellectual engagement, whereas Openness reflects engagement with fantasy, perception and aesthetics. Strong evidence regarding the predictive validity of this dissociation was provided by Kaufman et al. (2016) who demonstrated that Openness predicts creative achievement in the arts, whereas Intellect predicts creative achievement in the sciences (see also Kaufman et al., 2010). The BFAS was included to provide a sense of the level of creativity of our sample, from the perspective of personality structure.

2.3. Image acquisition

A 3-Tesla MR scanner with an 8-channel head coil (Discovery MR750, 22.0 software, GE Healthcare, Waukesha, WI) was used to acquire T1 anatomical volume images (0.86 x 0.86 x 1.0 mm voxels, number of slices = 26; repetition time [TR] = 2000 ms; echo time [TE] = 30 ms; flip angle [FA] = 70°; field of view [FOV] = 200 x 200 mm²; 64 x 64 matrix), positioned to cover the whole brain.

2.4. Statistical analysis

The data were analysed using Statistical Parametric Mapping (SPM12) (http://www.fil.ion.ucl.ac.uk/spm/software) implemented in Matlab (http://www.mathworks.com/products/matlab/). Before conducting the co-registration and normalization steps in SPM12, we manually reset the origin of our anatomical images (i.e., template space) using the Display function, with the anterior commissure as the reference point. Image registration was conducted using the Check Reg function and rigid-body registration applied accordingly. All images were segmented into grey matter, white matter, cerebrospinal fluid, skull, extra-skull, and air. Beginning with pre-processing, the specifications for the segmentation process were as follows: we maintained at SPM12 default values for channel bias regularization (0.0001), bias Full-Width Half-Maximum (FWHM, 60 mm cutoff), and bias correction. For grey matter we maintained default values for tissue probability map and Number of Gaussians (= 1), and selected Native + DARTEL to enable image generation for DARTEL registration. DARTEL is a template creation method that increases the accuracy of inter-subject alignment by modelling the shape of the brain using multiple parameters—in the form of three parameters per voxel. Warping and affine regularization were maintained at their default values. Smoothness (= 0) and sampling distance (= 3) were set to their default values. Field deformation was not applied. We ran DARTEL (by creating two channels for grey and white matter), following which images were spatially normalized to the Montreal Neurological Institute (MNI) brain template. This step generates smoothed, spatially normalised and Jacobian-scaled grey matter images in MNI space. Gaussian FWHM was set to 8 mm. For global normalization (the process whereby brains of different sizes and shapes are adjusted to enable group-level inferences) we selected relative masking with a threshold of 0.8. We opted for no global calculation or global normalization (the process whereby preprocessed data are scaled proportionally to the fraction of the brain volume accounted for by the represented grey matter). Subsequently, we conducted a multiple regression analysis in the GLM (General Linear Model) to capture the relation between regressors of interest and grey matter volume. Specifically, we regressed each of the seven scoring approaches (i.e., fluency, flexibility, originality, elaboration, snapshot creativity, novelty, and usefulness) onto grey matter volume—exploring both positive (weight = 1) and negative (−1) associations. We also entered sex, age, intracranial brain volume, verbal intelligence and working memory span as covariates in the analysis. Brain volume was calculated using the fslnums image calculator tool from FMRI Software Library (FSL) (Jenkinson et al., 2012). Of the six segmented tissues (described earlier), we added the grey and white matter tissues together for a brain image removed of cerebrospinal fluid, skull, and any air artifacts for each participant. A binary mask of this brain image was created to ensure that overlapping voxels, such as the boundary between grey and white matter, were not included multiple times in the brain volume calculation. We then used the fslnums tool to calculate the total volume of all voxels composing the binary mask of each participant, which is a measure of the brain volume. Each contrast produced a statistical parametric map consisting of voxels where the z-statistic
was significant at $p < .001$. Reported results survived voxel-level intensity threshold of $p < .05$ (corrected for multiple comparisons using the Bonferroni whole-brain family-wise error [FWE]).

### 2.5. Data and code availability statement

The data and code (i.e., scripts) used in this study are available upon direct request from the first author for secondary use of data (e.g., meta-analysis). The data-sharing procedure must comply with the requirements of Defence Research and Development Canada’s Human Research Ethics Committee. Specifically, in order to preserve the confidentiality of the participants, any publication arising from the data collected in this study will involve aggregate/group results only. No information which could be used to infer the identity of individual participants will be published. Toward that end, to preserve the anonymity of the participants any shared data will be paired with a personal identification number (PIN) rather than any personally identifying information.

## 3. Results

### 3.1. Behavioural

Average BFAS Intellect and Openness scores were 3.69 ($SD = 0.81$) and 3.59 ($SD = 0.58$) respectively. These values are comparable to those reported by DeYoung et al. (2007) derived from a community sample (Intellect = 3.70 ($SD = 0.68$), Openness = 3.74 ($SD = 0.61$)) as well as those derived from a university sample (Intellect = 3.39 ($SD = 0.67$), Openness = 3.52 ($SD = 0.64$)). This suggests that in terms of the BFAS aspects most relevant to creativity, the present sample is comparable to other community and university samples. In turn, Openness and Intellect were combined to compute the Openness/Intellect factor score ($M = 3.61$, $SD = 0.55$).

The means and standard deviations of the traditional, subjective, and definitional scores of the AUT are shown in Table 2. For fluency for which we had previous data based on a similar administration procedure (Vartanian et al., 2007, 2009), the average number of generated uses was comparable but somewhat lower than what had been observed before (8.62 ($SD = 2.98$) vs. 10.03 and 10.90). In turn, the correlations among the scores are shown in Table 3. As can be seen, most scoring approaches are correlated relatively strongly in the expected directions. Next, we sought to examine the correlations between each scoring approach and measures of intelligence and simple working memory span. Measures of simple working memory span were highly correlated: Word span and visual span ($r(40) = 0.51, p < .001$); word span and matrix span ($r(40) = 0.53, p < .001$); and visual span and matrix span ($r(40) = 0.79, p < .001$). Therefore, they were averaged and standardized into a single simple working memory span score. As can be seen, only elaboration and novelty were correlated with intelligence and simple working memory span (Table 4).

Next, we examined the correlations between BFAS aspects and measures of intelligence as well as working memory span. As expected (DeYoung et al., 2005), Intellect was correlated positively with word span ($r(36) = 0.51, p < .01$), matrix span ($r(36) = 0.43, p < .01$), visual span ($r(36) = 0.41, p < .05$), and Shipley-2 Vocabulary ($r(36) = 0.63, p < .001$), whereas Openness did not exhibit correlations with any of those variables ($p > .05$). Neither Intellect or Openness was correlated with Shipley-2 Abstract reasoning or Block design scores ($p > .05$). Openness/Intellect was not correlated with any of the seven scoring variables in the present study ($p > .05$).

There is an extensive literature in support of the serial order effect, according to which the originality of ideas is greater in later compared to earlier phase of divergent thinking production (Beaty and Silvia, 2012; Christensen et al., 1957; Johns et al., 2001; Phillips and Torrance, 1977; see also Gabora 2018, L. 2019). To explore that possibility we split the original uses generated on each item into those that had been generated in the first versus in the second half (chronologically). In accordance with the serial order effect, we found that there was a greater frequency of original responses in the second ($M = 7.05$, $SD = 4.64$) than in the first ($M = 3.10$, $SD = 2.30$) half, $t(41) = 6.99$, $p < .001$. We also examined the correlations between the frequency of generating original responses in the first versus the second half vis-à-vis our measures of intelligence and simple working memory span. Contrary to expectation, we found that the strength of the correlations did not vary as a function of the phase of divergent thinking production (see bottom of Table 4).

The present data also offered us the opportunity to explore the relative contributions of convergent and divergent thinking processes to performance on the AUT. Specifically, it has been suggested that convergent thought is more correctly defined as thought in which the relevant concepts are considered from conventional contexts, whereas divergent thought is defined as thought in which those concepts are considered from unconventional contexts (Gabora, 2018). Based on this framing, Gabora (2019) has argued that because in divergent thinking tasks (e.g., AUT) participants reflect upon an idea from unconventional contexts only after they have generated conventional responses, such tasks test for divergent thinking during the latter part of the task only. This proposal is consistent with findings that the later part of the task is typically when the most creative responses occur (Beaty and Silvia, 2012). In light of this, we reanalyzed the data by taking the total number of responses (i.e., sum of all responses—fluency) into account. Specifically, we divided each score (i.e., originality, flexibility, elaboration, snapshot creativity, novelty and usefulness) by fluency, and tested the prediction that a stronger pattern of correlations would emerge among scores that reflect the unconventionality of responses. This prediction was partly supported (see Table 5). For example, the strength of the correlation between snapshot creativity and novelty did increase when fluency was accounted for. However, perhaps the most striking aspect of this reanalysis was its ability to highlight the relationship between usefulness and elaboration to scores that reflect unconventionality, such as novelty and snapshot creativity. It appears that when raters judge the creativity of thought, the extent to which responses are developed and appear useful matters.

Finally, given our wide age range and the possibility that younger participants might have generated uses that were deemed more creative, we calculated the zero-order correlations between the participants’ age and fluency ($r = 0.08$, $p = .64$), originality ($r = 0.18$, $p = .26$), flexibility ($r = 0.09$, $p = .58$), elaboration ($r = 0.25$, $p = .12$), snapshot creativity ($r = 0.33$, $p = .04$), novelty ($r = 0.20$, $p = .21$), and usefulness ($r = -.02$, $p = .92$) scores. As can be seen, with the exception of snapshot creativity, none of the correlations reached statistical significance. Importantly, even in the case of snapshot creativity the correlation was positive, such that the uses generated by older participants were rated as more creative (as a whole).

### 3.2. Neural

Results from the multiple regression analysis revealed that novelty and usefulness were correlated negatively with regional GMV in the left
inferior temporal gyrus (ITG, BA 20) (Table 6). There was no correlation between fluency, originality, flexibility, snapshot creativity and elaboration with regional GMV. Next, to see whether there was a brain region where regional GMV covaried with divergent thinking performance across both novelty and usefulness, we conducted a conjunction analysis involving those two contrasts. The results demonstrated that regional GMV in the left ITG ($r = .87$, $x = -.56, y = -.29, z = -.21, k = 835, p = .039$) covaried with both novelty and usefulness scores (Fig. 2).

Finally, given that we had established the presence of a serial order effect, we examined whether there was a correlation between regional GMV and originality when the latter was split into the frequency of original responses generated in the first versus the second half of divergent thinking production. We found that there was no correlation between regional GMV and originality in either phase of divergent thinking production.

### 4. Discussion

Substantial work has already been conducted to explore the correlations between regional GMV and creativity, revealing a distributed set of regions that correlate negatively and/or positively with creativity across an assortment of measures (Table 1). The present study was conducted to explore whether the neural correlates of divergent thinking, as measured by the same task (i.e., AUT), would differ as a function of the specific way in which performance was scored. This hypothesis was motivated by two lines of work. First, previous behavioural studies have demonstrated that different scoring approaches to the AUT exhibit different patterns of correlations with measures of intelligence and executive functions (Benedek et al., 2012, 2014; Jauk et al., 2013, 2014; Zabelina et al., 2019). Behaviourally, we found the same here, as only elaboration and novelty were correlated with intelligence and simple working memory span. These findings suggest that each score might reflect a specific set of mental operations that underlie it, which in turn might recruit different neural systems. Second, we had discovered previously in an fMRI study that activation in the PCC predicts variations in performance as measured by fluency, flexibility, and snapshot creativity scores, but not in relation to originality, novelty, elaboration, or usefulness scores (Vartanian et al., 2019). Given that the relationship between brain function and divergent thinking performance varies as a function of scoring approaches, there existed the possibility that the same could be true if the focus were shifted to a measure of brain structure. Having statistically corrected for sex, age, intracranial brain volume, verbal intelligence and working memory span, our results demonstrated that usefulness and novelty were correlated negatively with regional GMV in the left ITG. There was no correlation between originality, fluency, flexibility, snapshot creativity or elaboration with regional GMV in the brain. Importantly, this pattern of findings emerged despite the presence of statistically significant correlations amongst these measures (see Table 3). This finding suggests that if AUT performance is scored in multiple ways, then there is merit in exploring the relationship between each scoring approach and regional GMV separately as that can reveal a more complete picture of the structural correlates of divergent thinking.

Our study is not the first to study the relationship between various ways of scoring divergent thinking performance and regional GMV. Notably, Jauk et al. (2015) investigated that correlation focusing on originality and fluency, and found that originality was correlated with re-

### Table 3

| Fluency | Originality | Flexibility | Elaboration | Snapshot | Novelty | Usefulness |
|---------|-------------|-------------|-------------|----------|---------|------------|
| .08     | −.11        | −.21        | −.18        |          |         |            |
| .16     | .06         | .02         | −.03        |          |         |            |
| .11     | −.07        | −.18        | −.21        |          |         |            |
| .34*    | .32*        | .37*        | .35*        |          |         |            |
| .34*    | .22         | .11         | .34*        |          |         |            |
| .14     | .11         | .03         | .05         |          |         |            |
| .14     | .11         | .03         | .06         |          |         |            |

Notes. AUT = Alternate Uses Task; WM = Simple working memory span; Origin. 1 = Originality first half of responses; Origin. 2 = Originality second half of responses. Vocabulary, Abstraction and Block Design scores were calculated based on the Shipley-2 (Shipley et al., 2009). Simple working memory span measure was a standardized average of verbal, visual, and matrix span (see Harrison et al., 2013; Vartanian et al., 2016). Adapted from O. Vartanian et al. (2019). * = $p < .05$.

### Table 4

| Vocabulary | Abstraction | Block Design | WM |
|------------|-------------|--------------|----|
| Fluency    | .08         | −.11         | −.18|
| Originality| .16         | .06          | .02 |
| Flexibility| .11         | −.07         | −.18|
| Elaboration| .34*        | .32*         | .37*|
| Snapshot   | .15         | .03          | −.04|
| Novelty    | .34*        | .22          | .11 |
| Usefulness | −.08        | .05          | .10 |
| Origin. 1  | .09         | .11          | .03 |
| Origin. 2  | .14         | .11          | .06 |

Notes. AUT = Alternate Uses Task; WM = Simple working memory span; Origin. 1 = Originality first half of responses; Origin. 2 = Originality second half of responses. Vocabulary, Abstraction and Block Design scores were calculated based on the Shipley-2 (Shipley et al., 2009). Simple working memory span measure was a standardized average of verbal, visual, and matrix span (see Harrison et al., 2013; Vartanian et al., 2016). Adapted from O. Vartanian et al. (2019). * = $p < .05$.

### Table 5

| Originality | Flexibility | Elaboration | Snapshot | Novelty | Usefulness |
|-------------|-------------|-------------|----------|---------|------------|
| −           | −.04        | −           |          |         |            |
| −.37*       | −.38*       | −           |          |         |            |
| −.31*       | .36*        | .95***      | −        |         |            |
| .04         | .70**       | .45*        | .43**    | −       |            |
| −.58**      | .27         | .85***      | .78***   | .31*    | −          |

Notes. AUT = Alternate Uses Task. For the scores here each measure was divided by sum of all responses (i.e., Fluency) (see text). * = $p < .05$; ** = $p < .01$; *** = $p < .001$. 

---

O. Vartanian, L. Smith and T.K. Lam et al.
NeuroImage 223 (2020) 117325
Table 6
Regional grey matter volume correlates of divergent thinking.

| Brain region | BA | T   | p     | Peak coordinates | Cluster |
|--------------|----|-----|-------|------------------|---------|
| Usefulness   | ITG| 20  | 6.29  | .015             | x −56   |
|              |    |     |       | y −29            | z −21   | 1184    |
| Novelty      | ITG| 20  | 5.87  | .024             | x −56   |
|              |    |     |       | y −29            | z −21   | 1055    |

Notes. BA = Brodmann Area; ITG = Inferior temporal gyrus, T = T-score; k = number of contiguous voxels; MNI = Montreal Neurological Institute. The p-value reflects a VBM threshold that survived Bonferroni whole-brain corrections for multiple comparisons (i.e., Family-Wise Error).

Fig. 1. The correlations between regional GMV in left inferior temporal gyrus (BA 20) and novelty and usefulness scores.

Notes. GMV = Grey matter volume.

Fig. 2. Regional GMV in left inferior temporal gyrus (BA 20) covaried with divergent thinking performance in relation to novelty and usefulness scoring approaches.

Notes: BA = Brodmann Area; GMV = Grey matter volume. The figure represents the result of a conjunction analysis involving the two contrasts for which a statistically significant negative correlations with regional GMV was found (i.e., novelty and usefulness). The region is overlaid on a single subject T1 image in SPM12 and reflects the coronal view. The bar represents the strength of the T-score. The T-score reflects a VBM threshold of p < .05 that survived a Bonferroni whole-brain correction for multiple comparisons (i.e., Family-Wise Error).

A much larger sample (n = 135) as well as the use of latent variable modelling to effectively account for measurement error in the observed variables. However, there are distinct differences between the two study designs. First, our scoring approaches extended beyond originality and fluency to include five other scores. Second, whereas all of our scores were derived from the same test (i.e., AUT), Jauk et al. assessed performance on the AUT as well as the instances task (which instructs people to generate instances of objects that meet certain criteria) to derive their creative potential scores for fluency and originality via structural equation modelling (SEM). The studies represent different approaches to investigating the correlation between regional GMV and various ways of scoring divergent thinking tests.

From an exploratory perspective, we also examined and found support for the serial order effect, according to which the originality of ideas is greater later compared to earlier in the course of divergent thinking production (e.g., Beatty and Silvia, 2012; Christensen et al., 1957; Johns et al., 2001; Phillips and Torrance, 1977). One reason for this effect could be that in the early phase of divergent thinking the most common strategy might involve the retrieval of easily available instances drawn from long-term memory—in line with the availability heuristic (Tversky and Kahneman, 1973). In turn, in the later phases of divergent thinking when such easily available uses have been exhausted, participants must engage in more effortful processes to retrieve and/or generate truly original uses. Support for this idea was provided in a study by Gilhooly, Floratos, Anthony, and Wynn (2007) who used a think-aloud paradigm to demonstrate that early uses were based on a strategy of retrieval of pre-known uses from long-term memory, whereas later responses relied on a number of strategies that draw more heavily on executive functions, including imagined disassembly of the target objects into components, and the reassembly of those components into novel objects, among others. Gilhooly et al.’s (2007) results do not
imply that retrieval from long-term does not play a role in the generation of original uses, but rather that relatively speaking, it can carry a greater role (in the absence of executive functions) in listing easily available uses in the early phase of divergent thinking performance. We examined the correlations between the frequency of generating original responses in the first versus the second half vis-à-vis our measures of intelligence and simple working memory span. Contrary to expectation, we did not find a difference in the strength of correlations between measures of intelligence or working memory span and the frequency of generating original responses in the first versus the second half of AUT performance (see Gilhooly et al., 2007). The same was true when we focused on regional GMV. Wang et al. (2017) recently investigated the neural correlates of the serial order effect using the electroencephalogram (EEG). Behaviourally, they probed the contribution of executive functions (i.e., updating, shifting, and inhibition; Miyake et al., 2000) to this phenomenon, and observed the serial order effect only for higher-order shifting. Furthermore, the EEG results revealed that it was only the lower-order individuals who exhibited stronger upper alpha (10–13 Hz) synchronization in left frontal areas during the early vs. late phases of AUT performance. These results suggest that the ability to shift increases the likelihood of observing the serial order effect, and that the ability to observe a correlation between regional GMV and divergent thinking scores might depend on the segregation of participants based on executive functions (e.g., inhibition). Future studies would be well-advised to also include a complete set of measures of executive functions (i.e., updating, shifting, inhibition) to probe this possibility.

Having said this, it is important to emphasize that the regulation of cognition in the service of creativity need not necessarily be a top-down process. For example, Martindale (1999) argued that switching between focused and defocused attention that enables creative people to consider ideas and concepts in novel ways is likely a bottom-up process that is triggered by the sensitivity of creative people to features of the problem space: When the problem space is ill-defined and ambiguous, attention becomes defocused, enabling the flexible consideration of ideas and concepts. In turn, when the problem space is well-defined and unambiguous, attention becomes focused on specific target ideas and concepts that require finetuning to completion. Along similar lines, considering a concept (e.g., brick) in a new context (e.g., needing a doorstep) also requires a shift in attention, and such a shift can arise spontaneously due to the sparse, distributed, content-addressable nature of memory (Gabora, 2017, 2018; L. 2019). Such shifts of attention, to the extent that they are spontaneous and reflect overlap in distributed representations, would appear to involve the default-mode network (DMN) rather than the executive control network (ECN).

Recently, in the context of her Model of Creativity and Attention (MOCA), Zabelina (2018) has distinguished between flexible vs. leaky attention. Flexible attention is evident early in the processing stream, and is driven by rapid focus, inhibition, and rapid shifting of attention. In contrast, leaky attention represents one’s propensity to notice information that other people see as irrelevant. Leaky attention is associated with weak cognitive flexibility and control, and with psychopathology-spectrum disorders. Zabelina (2018) has argued that whereas tests of divergent thinking (e.g., TTCT) measure flexible attention, measures of real-life creativity (e.g., CAQ) tap leaky attention. Indeed, in a series of studies Zabelina and colleagues have demonstrated that TTCT and CAQ scores have dissociable and theoretically-predictable relationships with measures of sensory gating based on both behavioural and neural measures (Zabelina et al., 2015, 2016). This work represents a fine example of matching specific creativity measures (i.e., TTCT vs. CAQ) with specific theoretical constructs related to attention (i.e., flexible vs. leaky attention). We advocate for the same approach in selecting the relevant scoring method for any given divergent thinking task. Specifically, we believe that the choice of which measures to select and how to score them should be theoretically guided, such that the specific measures and the associated scoring approaches are selected to reveal processes and mechanisms of interest that underlie creativity (see Abraham, 2016).

Related to the point above, the extent to which the scoring methods for divergent thinking tasks reflect theoretically-derived aspects of the cognitive architecture of creativity varies. For example, there is increasing convergence in the cognitive neuroscience literature that two large-scale systems underlie performance in divergent thinking tasks (for review, see Beatty et al., 2016). Specifically, the DMN appears to be the engine for the generation of novel ideas, whereas the ECN appears to be the system that exerts cognitive control over generated ideas to ensure that the output meets task demands. Importantly, the DMN and the ECN exhibit high levels of functional connectivity and temporally-variable involvement in the course of divergent thinking, suggesting that they act in concert to guide creative cognition. However, because we typically assess divergent thinking performance based on final output, at present it is debatable whether the contributions of the DMN and/or the ECN can be parsed out based on any scoring approach. This problem is amplified when we consider the specific cognitive sub-processes that contribute to DMN or ECN activity during divergent thinking. For example, from a theoretical perspective, semantic memory has historically been seen to play an indispensable role in the associative processes that support creativity (Mednick, 1962). More recent work has highlighted the contribution of episodic memory to creative cognition (see Madore et al., 2015). It would appear that semantic and episodic memory contribute to the generation of novelty, and yet we do not have theoretically-derived scoring procedures that can isolate their respective contributions with sufficient granularity. As noted by Reter-Palmon et al. (2019), our domain is in need of stronger theoretical justification with respect to measurement choices so that ultimately we can make inferences about the cognitive processes of interest based on specific scores.

Our results demonstrated that regional GMV in the left ITG (BA 20) was correlated negatively with usefulness and novelty scores. Substantial neuroimaging and patient data suggest that this region is part of the brain’s core semantic system, and is involved in concept retrieval and integration (Badre and Wagner, 2007; Thompson-Schill, 2003). In addition, a large-scale meta-analysis meant to isolate the semantic system in the brain revealed that the system can be fractionated into seven distinct subsystems, each with its specific functions (Binder et al., 2009). One of those seven subsystems overlaps with several regions in the lateral and ventral left temporal lobe, including most of the middle temporal gyrus and portions of the ITG, fusiform gyrus, and the parahippocampus. Binder et al. (2009) suggested this subsystem appears to be heteromodal in function, and is involved in supramodal integration and concept retrieval. As elaborated further by Binder and Desai (2011), the combination of the neuroimaging and patient data “offer compelling evidence for high-level convergence zones in the inferior parietal, lateral temporal, and ventral temporal cortex. These regions are far from primary sensory and motor cortices and appear to be involved in processing general rather than modality-specific semantic information” (p. 530). Understood as such, it is perhaps not too difficult to see that ITG could be involved in usefulness and novelty scores, given that both scores impose varying degrees of requirement on concept retrieval and integration involving the semantic system.

Indeed, the left temporal lobe has been implicated repeatedly in neuroimaging studies of creativity, although the observation of the involvement of the left ITG in particular is novel (see Boccia et al., 2015; Gonen-Yaacovi et al., 2013; Wu et al., 2015). We had earlier investigated the relationship between cortical volume and thickness and variation in scores on the Big Five personality factor of Openness/Intellect (i.e., openness to experience) — defined as “the breadth, depth, originality, and complexity of an individual’s experiential life” (John et al., 2008, p. 120). We found that Intellect was uncorrelated with cortical thickness or volume, whereas Openness was correlated negatively with cortical thickness and volume in six regions of the brain including the left middle temporal gyrus (BA 21) and the left superior temporal gyrus (BA 41) (Vartanian et al., 2018b). This pattern of results is informative because unlike studies of intelligence that typically show a positive correlation between measures of cognitive ability and cortical thickness
and/or volume (e.g., Draganski et al., 2004; Haier et al., 2005), a recent review of structural studies of creativity has shown that the brains of individuals show patterns of increased as well as decreased cortical volume/thickness in relation to test scores (Jung et al., 2013). Furthermore, addressing the question of why it is that reduced rather than increased regional GMV in left ITG exhibits covariation with novelty and usefulness likely requires both developmental and patient data to uncover its specific functional contribution to creative cognition as measured by usefulness and novelty scores.

Developmental and patient data are also relevant to our understanding of why, despite their negative correlation behaviourally, novelty and usefulness were both found to exhibit a negative correlation with regional GMV in ITG. For example, using a longitudinal design, Shaw et al. (2006) were able to demonstrate that whereas in early childhood there was a negative correlation between intelligence and cortical thickness, in late childhood and beyond there was a marked shift to a positive correlation between intelligence and cortical thickness. Importantly, this developmental shift was particularly salient in the frontal lobes which are implicated heavily in intelligent behaviour in adults. At the moment we have no comparable data regarding the relationship between creativity and cortical maturation (Vartanian, 2020). As a result, it is difficult to interpret the meaning of the direction of the correlation between any creativity score and measures of brain structure such as cortical thickness or volume. It could be that regional GMV in any given region that contributes to creativity may be associated with the facilitation or inhibition of different capacities at different time points throughout life. To make inferences regarding directionality possible, developmental and patient data are needed to shed light on the trajectory of cortical maturation in relation to creativity—measured in different ways.

Here it is also important to ask why it is that regional GMV was related to novelty and usefulness, but not to the other five indices. One reason might be differences in the specific cognitive processes that underlie each of those scores, and the extent to which those processes draw on the ITG in terms of concept retrieval and integration demands. Novelty was defined in terms of uncommonness, whereas usefulness was defined in terms of feasibility. For a use to qualify as novel or useful, the participant likely had to retrieve concepts from long-term memory and integrate them in uncommon or meaningful (i.e., feasible) ways, respectively. In turn, fluency, flexibility, and elaboration would appear to place greater press on the amount and breadth of concept retrieval rather than their integration per se. For originality, we opted for the uniqueness scoring method (Wallach and Kogan, 1965), according to which a value of “1” was given to each response that was not generated by any other participant in this sample, and a value of “0” to any response that was generated by at least one other participant in this sample. As noted by Silvia et al. (2008), uniqueness scoring is strongly influenced by the relative performance of others in the sample, and as such novelty scores here may be a more generalizable representation of uncommon responding than originality, despite their high correlation. Finally, because snapshot creativity scores involve giving a single score to the entire set of uses generated by a participant (Silvia et al., 2008), they might not be a true representation of the most creative thoughts a participant might have had, as would have been the case with other approaches such as the subjective top-scoring method (see Benedek et al., 2013). In summary, we believe that as conceptualized here, novelty and usefulness were the two scoring approaches that drew on the ITG in terms of concept retrieval and integration demands to a greater extent.

Our study has some limitations. First, our investigation was based on a single task (i.e., AUT). As such, whether the neural correlates of other divergent thinking tasks prove equally sensitive to these scoring approaches remains to be seen. From a more theoretical perspective, we also know that creativity requires both convergent and divergent thinking (e.g., Beersma and Dreu, 2005; Gibson et al., 2009; Kerr and Murphy, 2004). It is likely that the AUT as a task, or any of the scoring approaches in relation to the AUT, will have limited utility for fully highlighting the contributions of both convergent as well as divergent thinking to creative cognition. We tried to ameliorate this problem by including scoring approaches that likely vary in the extent to which they draw on divergent vs. convergent processes (e.g., fluency vs. snapshot creativity or originality), but this problem cannot be addressed in the absence of using a menu of tasks that load differentially on convergent and divergent thinking. Second, although sample sizes for VBM studies of creativity have varied greatly (range 18–366, Table 1), our sample size was small. Furthermore, DARTEL uses the spatial intensity distribution of the structural MRIs of the specific subjects under consideration to create a group template, which is in turn used to create final normalized tissue maps for each subject in the group (Michael et al., 2016). For these reasons, replications of our findings will be necessary to determine their reliability. Third, although as a field we are gaining traction on the relationship between various scoring approaches and constructs such as intelligence and executive functions (Benedek et al., 2012, 2014; Jauk et al., 2013, 2014; Zabelina et al., 2019), more work is necessary to develop a fully componential model involving the contributions of key cognitive ability and personality factors to each scoring approach. Fourth, although the BFAS scores of our sample were within the ranges of what has been measured before in both community and university samples (DeYoung et al., 2007), the fluency scores observed here were comparable to but somewhat lower than what had been observed in our previous studies (Vartanian et al., 2007, 2009). As such, it will be important to determine whether the results obtained in the present study will extend to samples that exhibit higher degrees of performance on divergent thinking tasks, including the AUT. Nevertheless, it is our hope that our preliminary work presented here will contribute toward improved approaches to creativity assessment—in particular the scoring of divergent thinking tasks and their associations with brain structure and function.

Acknowledgements

This work was supported by funding from the Department of National Defence (DND) to a work breakdown element (WBE) under the direction of the first author (OKR04). We acknowledge the contributions of Emma Rice, Sarah Forbes and Jenna Crocker to this research.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117325.

References

Abdulla, A.M., Paeck, S.I., Cramond, B., Runco, M.A., 2020. Problem finding and creativity: a meta-analytic review. Psychol. Aesth., Creat. Arts.

Abraham, A., 2016. Commentary: creativity and memory: effects of an episodic-specificity induction on divergent thinking. Front. Psychol. 7. doi:10.3389/fpsyg.2016.00824, Article 824.

Amabile, T.M., 1982. Social psychology of creativity: a consensual assessment technique. J. Pers. Soc. Psychol. 43, 357–376. https://doi.org/10.1037/0022-3514.43.5.997.

Arkin, C., Prayvinida, E., Pfefier, C.W., Zeng, T., Loui, P., 2019. Gray matter correlates of creativity in musical improvisation. Front. Hum. Neurosci. 13. doi:10.3389/fnhum.2019.00169, Article 169.

Badde, D., Wagner, A.D., 2007. Left ventrolateral prefrontal cortex and the cognitive control of memory. Neupysysiology 45, 2883–2901. https://doi.org/10.1016/j.neupysy.2007.06.015.

Barbot, B., Reiter-Palmon, R., 2019. Creativity assessment: pitfalls, solutions, and standards. Psychol. Aesthet. Creat. Arts. 13, 131–132. http://dx.doi.org/10.1037/aest0000251.

Beatty, R.E., Benedek, M., Silvia, P.J., Schacter, D.L., 2016. Creative cognition and brain network dynamics. Trends Cogn. Sci. (Regul. Ed.) 20, 87–95. https://doi.org/10.1016/j.tics.2015.10.004.

Beatty, R.E., Kenet, Y.N., Christensen, A.P., Rosenberg, M.D., Benedek, M., Chen, Q., Fink, A., Qiu, J., Kwapiel, T.R., Kane, M.J., Silvia, P.J., 2018. Robust prediction of individual creative ability from brain functional connectivity. Proc. Natl. Acad. Sci. USA 115, 1087–1092. https://doi.org/10.1073/pnas.1713321115.

Beatty, R.E., Silvia, P.J., 2012. Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. Psychol. Aesthet. Creat. Arts. 6, 309–319. http://dx.doi.org/10.1037/a0029171.
