Neurological Divide: An fMRI Study of Prose and Code Writing

Ryan Krueger
University of Michigan
ryankrue@umich.edu

Yu Huang
University of Michigan
yhhy@umich.edu

Xinyu Liu
Georgia Institute of Technology
xinyuliu@umich.edu

Tyler Santander
UC Santa Barbara
t.santander@psych.ucsb.edu

Westley Weimer
University of Michigan
weimerw@umich.edu

Kevin Leach
University of Michigan
kjleach@umich.edu

ABSTRACT

Software engineering involves writing new code or editing existing code. Recent efforts have investigated the neural processes associated with reading and comprehending code — however, we lack a thorough understanding of the human cognitive processes underlying code writing. While prose reading and writing have been studied thoroughly, that same scrutiny has not been applied to code writing. In this paper, we leverage functional brain imaging to investigate neural representations of code writing in comparison to prose writing. We present the first human study in which participants wrote code and prose while undergoing a functional magnetic resonance imaging (fMRI) brain scan, making use of a full-sized fMRI-safe QWERTY keyboard.

We find that code writing and prose writing are significantly dissimilar neural tasks. While prose writing entails significant left hemisphere activity associated with language, code writing involves more activations of the right hemisphere, including regions associated with attention control, working memory, planning and spatial cognition. These findings are unlike existing work in which code and prose comprehension were studied. By contrast, we present the first evidence suggesting that code and prose *writing* are quite dissimilar at the neural level.

KEYWORDS

medical imaging, spatial, memory, attention, synthesis, keyboard

ACM Reference Format:
Ryan Krueger, Yu Huang, Xinyu Liu, Tyler Santander, Westley Weimer, and Kevin Leach. 2020. Neurological Divide: An fMRI Study of Prose and Code Writing. In *42nd International Conference on Software Engineering (ICSE '20), May 23–29, 2020, Seoul, Republic of Korea.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3377811.3380348

ICSE '20, May 23–29, 2020, Seoul, Republic of Korea

ACM ISBN 978-1-4503-7121-6/20/05…$15.00
https://doi.org/10.1145/3377811.3380348

1 INTRODUCTION

Writing code is a crucial activity in software engineering. With software-related innovations driving a $3.8 trillion global IT market [24] and demand for university computer science courses outstripping the supply of professors [107], the value of developing and maintaining software is increasing rapidly. This importance is already reflected at an industrial scale, with Fortune 500 companies, such as Amazon and AT&T, committing massive resources to retrain up to half of their workforce in programming-intensive areas [20, 27]. Despite this increasing prevalence of software and demand for skilled programmers, we rely on traditional survey instruments and self-reporting, rather than an understanding of fundamental human brain function, when developing methods to support, improve, teach and evaluate code writing and editing. We present findings from the first study to use medical imaging to investigate the cognitive processes underlying the writing of code.

Problem. There is a significant body of work studying the psychology of programming, ranging from the cognitive prerequisites of programming [80] to entire theories of the coding process [13], but this research has relied largely on observational evidence. Recent advances in medical imaging, particularly functional magnetic resonance imaging (fMRI), have improved researchers' ability to measure brain activity associated with various cognitive processes. As a non-invasive, *in vivo* technique, fMRI is an effective tool for clinical researchers studying brain function [1, 42, 110] and the effects of various treatments [76, 102, 113], as well as for psychology researchers mapping brain areas in activities as diverse as musical performance [65] and food cravings [86]. Findings using medical imaging have successfully transitioned to guiding behavioral and developmental improvement in domains like mathematics [30] and education [87]. While medical imaging studies are still new in computer science, software engineering researchers have used fMRI to help understand tasks like code comprehension [103], code review [39] and data structure manipulation [52]. Imaging advances hold out the promise of helping computer science as they have helped other fields: from understanding expertise [4, 67, 98] to retraining an aging workforce [19, 74] to guiding pedagogy [5, 97] to augmenting unreliable self-reporting [66, 90].

Challenge and Insight. While there have been fMRI studies of code *reading* (e.g., [39, 103]) and *non-fMRI* studies of code writing (e.g., [12, 68]), to the best of our knowledge there are no previous fMRI studies of code writing. We attribute this to two challenges: physics and design. First, normal keyboards cannot be safely placed or accurately read near magnetic resonance scanners. They interfere

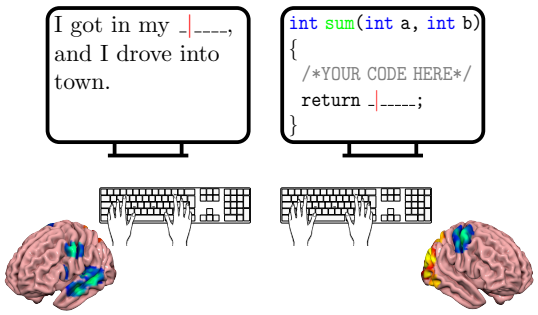


Figure 1: We investigate the relationship between prose and code writing using functional brain imaging. Experimental controls systematically vary content (code vs. prose) and size (fill-in-the-blank vs. long response). Do code and prose writing exhibit the same patterns of neural activity?

with the fMRI measurements and the fMRI interferes with keyboard reporting. Second, imaging studies require carefully-controlled experiments, and no high-level design for a code writing contrast has been proposed (cf. Behroozi *et al.*'s contrast of whiteboard interview questions with pencil-and-paper versions [7], which changes the modality but uses identical tasks, or Huang *et al.*'s contrast of data structure problems with mental rotation problems [52], which changes the task but not the modality). We combine two corresponding insights to overcome these challenges. First, we propose to employ a bespoke keyboard that moves all metal and control logic to a separate room. Second, we propose a two-by-two contrast setup: code vs. prose writing and fill-in-the-blank vs. long response (informally, single-word production vs. longer creativity).

Our use of prose writing as a baseline grounds our experiment and clarifies our results. Prose writing is a well-studied activity in psychology [6, 26, 47, 57, 64], and medical imaging has aided understanding of its underlying cognitive processes. For example, fMRI studies have provided insights into brain areas associated with prose writing [73] and the specificity of such regions across different prose writing tasks [89], in addition to addressing neural correlates of the roles of expertise [95] and creativity [101]. The contrast between code and prose writing in our experiment illuminates their differences and similarities at a neurological level.

Experiment. We conducted a human study in which 30 participants performed prose and code writing tasks in an fMRI scanner (Figure 1). Participants completed two types of tasks: fill-in-the-blank (FITB) and long response (LR). FITB tasks presented either a sentence or program containing a blank space, requiring the participant to provide the missing word or code snippet. In LR tasks, participants wrote prose or code from scratch to answer an open-ended question or meet a program specification.

Results. Our primary finding is that code writing and prose writing feature significantly different patterns of neural activity, particularly in parts of the brain associated with attention control, working memory, and spatial cognition. While prose writing involves activation in canonical areas associated with language, code writing involves a very different set of right-lateralized regions associated with attention, memory, planning and spatial ability. Our experiment provides the first evidence of significant neural differences between *prose writing* (which is neurally similar to natural language) and *code writing* (which, we find, is *not*).

The contributions of this paper are as follows:

- An fMRI study of 30 participants comparing code writing to prose writing. To the best of our knowledge, this is the first fMRI study to feature keyboard code writing. Our experimental design contrasts code, prose, fill-in-the-blank and long-response questions.
- A mathematical analysis of the results. After mitigating noise and correcting for false discovery rate ($q < 0.05$), we find that general code and prose writing feature distinct patterns of neural activity ($2.4 \leq t \leq 6.2$) related to attention, working memory and spatial cognition. For long-response writing questions, we find the clearest distinction we are aware of in the literature ($-7.0 \leq t \leq -3.1$ and $3.5 \leq t \leq 5.8$) between code (attention, memory, planning and spatial ability) and prose (language, letters and words).
- For replication and reproducible research, we make available our materials and methods on our project website.[1] These include our corpus of stimuli; our de-identified medical imaging data; our method for adapting a 101-key QWERTY USB keyboard for the fMRI environment; and a configurable program for stimuli presentation, editing and data collection.

2 BACKGROUND

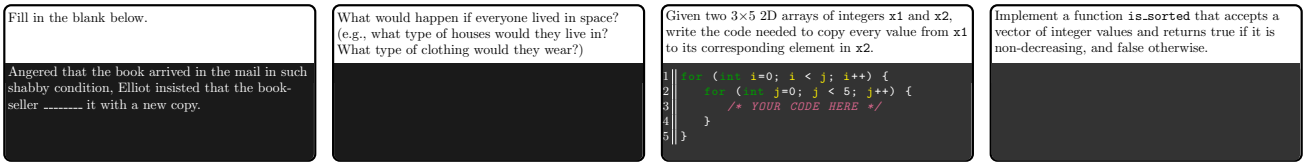
In this section, we summarize how medical imaging can uncover neural correlates of cognitive processes, previous non-writing studies of code, and previous non-imaging studies of writing.

2.1 Medical Imaging and Software Engineering

Functional magnetic resonance imaging (fMRI) is a measurement technique that has provided many neuroscience and cognitive science insights. Of the non-invasive, *in vivo* functional neuroimaging approaches available, fMRI has emerged as the most popular over the last 30 years, appearing in over 100,000 studies [44]. fMRI relies on the *hemodynamic response* or rapid change in blood flow to deliver metabolites (e.g., oxygen, glucose) to neuronal tissues. Oxygenated and deoxygenated blood cells vary in their magnetic properties, allowing fMRI to indirectly measure these changes via the application of magnetic fields. Using the energy released to locate these blood cells, fMRI calculates the *blood-oxygen level dependent* (BOLD) signal, defined as the ratio of oxygenated to deoxygenated hemoglobin.

The use of medical imaging to study software engineering is relatively new. There have been fewer than a dozen studies at major conferences that use fMRI to investigate software engineering [22, 34, 36, 39, 52, 53, 75, 84, 84, 103, 104]. Other medical imaging modalities have been used to examine software engineering (e.g., fNIRS [36, 75] or EEG [25, 63]) but Huang *et al.* demonstrated the importance of fMRI's penetrating power and spatial resolution when studying computing tasks [52, Sec. V-D]. All previous fMRI studies of software engineering have focused on reading or reviewing code; we make use of fMRI to study code writing.

[1] https://web.eecs.umich.edu/~weimerw/fmri.html



(a) Prose Fill-in-the-Blank **(b) Prose Long Response** **(c) Code Fill-in-the-Blank** **(d) Code Long Response**

Figure 2: Illustrative examples of stimuli shown to participants. We investigated four categories of stimuli covering code and prose in fill-in-the-blank and long response scenarios.

fMRI's reliance on the hemodynamic response and use of magnets both restrict study design and ecological validity. First, the temporal structure of the hemodynamic response limits measurements to small (typically less than one minute) windows. Second, the fMRI scanner and magnetic fields preclude proximate metallic or electronic devices: previous studies used a small, fixed number of "multiple choice" responses (e.g., selected by pneumatic button press or similar device). In this work, we lift this second restriction by making use of a full keyboard.

2.2 Writing and Cognition

There is strong interest from both academia and industry in improving programmers' ability to write code [40, 60, 68, 70]. While the success of a handful of projects (e.g., Scratch [93]) hints at the opportunity present in understanding the mental processes associated with writing code, modern research is limited by our lack of a foundational, neurobiologically-grounded understanding [52].

Researchers since the 1950's have sought to understand the psychology of programmers, but have largely relied on observational data (e.g., [14]). These efforts have ranged from studies of expertise [46, 78, 114, 117], to entire theories of the coding process [13], to how programmers use the Internet [12]. Pea and Kurland, who conducted research in the 1980's on the cognitive prerequisites and effects of computer programming, emphasized the need to understand programmers' psychology given the rising importance of computer literacy [80, 81]. Despite the massive growth of software engineering since then, Pea and Kurland's call has not been answered with a grounded neurological understanding.

There is also a significant body of research outside of computer science to study the cognitive processes of prose writing (see Berninger and Winn [10] for a survey). Unlike code, prose writing has been studied using medical imaging to establish a more objective understanding. The findings from these medical imaging prose writing studies, such as the brain regions associated with prose writing and the specialization of such regions [9, 73, 88, 89, 95, 101], have in turn successfully informed subsequent research in pedagogy [10, 57, 87].

Previous work simultaneously studying code and prose writing has focused on non-imaging uses of one to aid in the instruction of the other. More explicitly, research in storytelling has used programming as a means to improve children's prose writing [17], and vice versa [16, 56]. In either direction, researchers reported similarity in the processes of code and prose writing, such as their sequence, structure, and object-oriented nature [35]. However, these qualitative findings have not been substantiated by medical imaging.

2.3 Motivation: Imaging of Code Writing

Medical imaging has made remarkable contributions in guiding behavioral enhancement and development in different domains, such as mathematics [30] and education [87]. For instance, cognitive understanding of numeracy has inspired researchers to use different measures to predict individual differences in mathematical development and achievement [11, 48, 55, 62]. Based on medical imaging research in music training, researchers successfully developed interventions to enhance executive functioning and working memory in older adults [15]. Similarly, imaging findings in reading-related brain activities made it possible to design interventions to improve reading skills over time in dyslexic children [71, 106].

Surveyed educators largely believe understanding the brain is important to the design and delivery of teaching [87]. Berninger and Winn found that integration of neuroscience and learning science may promote educational evolution [10]. Dahlin *et al.* found that training can transfer between two tasks that engage overlapping processing components and brain regions [29]. Specifically, the neuroimaging findings of the role of working memory in prose writing [9, 94] have led to a series of instructional intervention studies showing writing problems can be improved [51, 58, 112]. Inspired by research on prose writing and other domains using medical imaging, we believe similar benefits for code writing may be available.

To the best of our knowledge, there are no previous imaging studies of code writing in general, nor studies comparing code and prose writing in particular; we return to non-imaging code writing research, including software psychology studies, in Section 7.

3 EXPERIMENTAL SETUP AND METHODS

We present a human study in which 30 participants underwent an fMRI scan while completing prose and code writing tasks. We discuss (1) the makeup and recruitment of our participant cohort, (2) how we developed our task materials, (3) the experimental protocol, (4) our method for collecting fMRI data, and (5) the construction of an fMRI-safe keyboard that enabled participants to freely write text and code during an fMRI scan.

3.1 Participant Demographics and Recruitment

We recruited 30 undergraduate and graduate computer science students at the University of Michigan. The protocol was approved by the University's IRB (HUM00138634). Table 1 summarizes demographic information for this cohort. Students who had completed coursework in data structures and who could safely undergo an MRI scan were eligible to participate. All participants were native

Table 1: Demographics of the participants in our study.

Demographic Variables	# Participants	
Sex	Male	20
	Female	10
Gender	Men	20
	Women	9
	Fluid	1
Degree Pursuing	Undergraduate	27
	Graduate	3

English speakers, right-handed, and had normal or corrected-to-normal vision. Each participant was offered a $75 cash incentive and a 3D model of their brain upon completion.

When participants elected to participate in the study, we collected basic demographic data (sex, gender, age, cumulative GPA, and years of experience) and socioeconomic status (SES) data. In addition, each participant completed three standard psychological measurement surveys: Positive and Negative Affect Scale (PANAS, emotional health), Autism Spectrum Disorder (ASD), and Need for Cognition (NFC, inclination for effortful cognition). Finally, we administered a short programming quiz to assess basic C/C++ programming skills.

Although we conducted a correlation analysis between these demographic and psychological measures and brain activities, none survived a strict false discovery rate correction ($q < 0.05$). We claim no significant demographic or attitudinal correlation with code or prose writing in our study. In the remainder of this paper, we thus treat our participants as a whole, rather than considering any subpopulation analyses.

3.2 Participant Tasks

Participants underwent an fMRI scan during which they completed a sequence of tasks associated with code and prose writing. Participants were shown a sequence of sentences or code snippets and asked to type a response while inside the MRI machine. We divided tasks into Fill-in-the-Blank (FITB) and Long-Response (LR) activities. In FITB, participants were shown a nearly-completed sentence or code snippet and had 30 seconds to type a short word or expression that they thought best completed the sentence or snippet. In LR, participants had 60 seconds to write a complete response to a high-level task or question. Participant completed four categories of tasks, each lasting 20 minutes: (1) 17 FITB Prose tasks, (2) 9 LR Prose tasks, (3) 17 FITB Code tasks, and (4) 9 LR Code tasks. Examples of stimuli under each of these categories are shown in Figure 2.

Code Tasks. We developed a corpus of code stimuli by adapting tasks from Turing's Craft [3], a library of short programming exercises used in web teaching evaluations [2], each with prompts and example correct solutions. For the FITB Code tasks, we selected a set of 17 prompt-answer pairs, and replaced a random portion of the solution with a blank line. Participants were asked to fill in that blank line. For the LR Code tasks, we selected a set of 9 prompts that our pilot study suggested as answerable within 60 seconds.

Prose Tasks. For controlled experimentation and to admit a contrast-based analysis, we selected prose stimuli that were analogous to the code stimuli. As prose writing fMRI studies have revealed differences in brain activation based on writing content [89, 101], we carefully developed our prose writing stimuli. First, we used a set of non-math analogies that have been shown to be useful in the teaching of mathematics [28, 99] to develop a list of terms associated with quantitative reasoning. Synonyms of these words were added to expand the search space. To generate Prose FITB prompts, we first matched the list of search words to a set of Scholastic Assessment Test (SAT) fill-in-the-blank questions and chose 17 such matches. We then replaced the blanks used in the original SAT prompt with the appropriate words from the SAT answer, selecting easier synonyms when our pilot study revealed that they might not be accessible to a wide population. We replaced the search word found in the prompt with a blank line; participants were asked to fill in that blank line. Our Prose LR prompts were generated by matching search words with a set of English as a Second Language (ESL) long response prompts and choosing 9 matching prompts.

29 out of the 30 recruited participants supplied valid inputs for the tasks. Per task, the 29 participants provided a maximum of 82 keystrokes (mean: 13) for FITB prose and 116 keystrokes (mean: 36) for FITB code. For the LR tasks, we collected a maximum of 435 keystrokes (mean: 258) for prose and 244 keystrokes (mean: 121) for code. In general, the FITB tasks required fewer keystrokes to complete; participants had twice the time to complete the LR tasks. We observed that participants were able to write multiple complete sentences for prose tasks and to complete variable declarations, loops, and function calls in the time allotted.

3.3 Experimental Protocol

In this subsection, we provide details on the process that participants completed before and during their fMRI scans. During a two-hour session, we collected informed consent and safety screening information. Participants cleared to participate were given a coding quiz and psychological surveys. Participants were then shown a brief training video about the task before entering the scanner. Each machine session began with a high-resolution anatomical scan during which participants were given a text editor interface and were instructed to practice typing on the keyboard while lying inside the bore of the machine (shown in Figure 3). This practice typing was not recorded. Participants then completed four task blocks associated with code and prose writing: Prose FITB, Prose LR, Code FITB, and Code LR. To mitigate training and fatigue effects, we randomized both the category order and the task order. A fixation cross was presented between each question for a random 2–10s duration to provide a brief rest and settle brain activity.

3.4 fMRI Data Acquisition

MRI data were acquired with protocols ensuring high spatial and high temporal resolution. We summarize the details (e.g., for the purposes of replication and meta-analysis), but generally attest that the scanning measurement hardware and steps align with contemporary best practices [39, 52, 103]. All scans were conducted on a 3T General Electric MR750 scanner with a 32-channel head coil at the University of Michigan Functional MRI Laboratory. First,

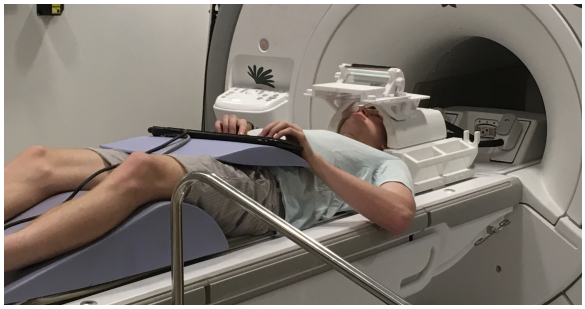


Figure 3: Typing in the fMRI machine. During a scan, the participant would be placed further in the bore of the machine, but the keyboard and visual interface remain as shown.

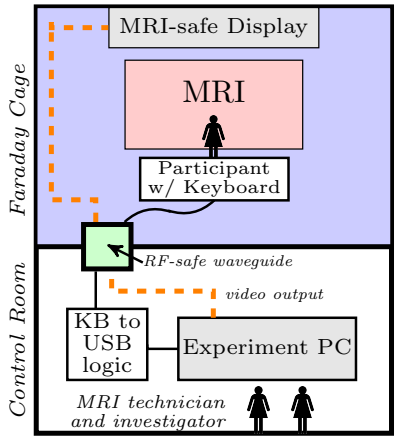


Figure 4: Illustration of fMRI writing setup. The participant lies in the bore of the fMRI machine, and the keyboard's cables are connected through an RF- and MRI-safe waveguide. The waveguide connects to the control room, where we attach the keyboard logic to an experiment PC displaying our editing environment. The video output of the experiment PC feeds through the waveguide to connect to an MRI-safe monitor which can be seen by the participant in the MRI bore via mirror projection.

high-resolution anatomical scans were collected with a T_1-weighted spoiled gradient recall (SPGR) sequence (TR = 2300.80 ms, TE = 24 ms, TI = 975 ms, $FA = 8^\circ$; 208 slices, 1 mm thickness). An estimate of magnetic field homogeneity was then acquired using a spin-echo fieldmap (TR = 7400 ms, TE = 80 ms; 2.4 mm slice thickness). All four subsequent task runs employed a T_2^*-weighted multiband echo planar imaging sequence (TR = 800 ms, TE = 30 ms, $FA = 52^\circ$; acceleration factor = 6) with whole-brain coverage over 60 slices (2.4 mm^3 isotropic voxels, or three-dimensional pixels).

3.5 fMRI-Safe Keyboard and Editing

Because the fMRI machine involves an extremely powerful magnet and very strong electromagnetic fields, typical electronic devices cannot be used safely nearby. For example, a traditional USB keyboard will not function in the MRI machine because it will induce current on the USB cable, causing erratic keystroke signals or unpredictable behavior. Moreover, large metal masses within the MRI's magnetic field can cause disastrous signal noise and ruin brain images (and also pose fire and collision hazards). Previous fMRI studies of software engineering all employed special hand-held button-press devices for selecting among a small, fixed set of choices (e.g., [39, 52, 103]). These devices do not meet the requirements for code writing.

In this work, we adapted a 101-key QWERTY USB keyboard for the fMRI environment. All control logic and metal are removed from the keyboard, and moving metallic pieces are replaced with 3D-printed (plastic) equivalents. Briefly, each individual key is attached to its own shielded wire that extends 30 feet to provide adequate distance from the core of the MRI machine. The wires were fed through an RF-safe waveguide to the fMRI control room, where a custom-built device reads the state of each key and outputs a standard USB signal. Because no control logic was present near the MRI machine and keystrokes were processed from the control room, we eliminated issues caused by electromagnetic interference. In addition, most fMRI studies use sequences of static pre-rendered stimuli controlled by software (e.g., E-Prime [91]) to record responses. We instead employed a more indicative dynamic editor environment, including syntax highlighting, available in our replication materials. We found this organization, illustrated in Figure 4, to work well, although imperfectly-printed plastic pieces caused occasionally-duplicated keystrokes for two participants.

While researchers have considered the problem for piano keyboards [109], keyboards with no screen (and thus no back-and-forth editing, e.g., [92]) or significant restrictions on which keys could be pressed (e.g., [50]), to the best of our knowledge, the closest related fMRI-keyboard work tends to be about a decade old (e.g., [54, 92]). Existing work has primarily focused on experimental design that reduces signal noise affecting brain scan quality. Our custom-built keyboard introduces negligible signal noise on the fMRI brain scan, but also supports our additional use case for live editing during scanning. We make our engineering notes on our successful approach (and failed attempts) available as part of our replication materials.

4 ANALYSIS APPROACH

Care must be taken when analyzing fMRI results to both mitigate noise and also to avoid false positive correlations [8]. Informally, we follow a three step process: preprocess the data to account for noise, analyze individual participants, and compare between participants. Our approach follows the state-of-the-art in medical imaging (both for cognitive neuroscience in general and for software engineering in particular, e.g., [39, 52, 103]). We present our results in Section 5; the remainder of this section summarizes our analysis for replication and comparison purposes.

Statistical analysis of fMRI data is inherently multi-level. The data first require extensive *preprocessing* to remove various sources of systematic noise (e.g., due to head motion or inhomogeneities in the magnetic field). An additional goal of this procedure is to align all individual participant brains with a standard anatomical template — this allows for inter-participant comparison and localization of signals to specific brain structures. Following preprocessing, each participant's data are submitted to a *first-level*, fixed effects general linear model (GLM). Here, voxel timeseries are modeled against an

idealized timeseries, given the canonical hemodynamic response function and the occurrence of each *event* (i.e., stimulus) over the course of the scan. This yields a set of *beta images* that describe how sensitive each voxel is to the conditions of our experiment. Finally, the beta images for each participant are combined in a *second-level*, random effects GLM, which yields average maps of brain activity when *contrasting* one condition versus another (e.g., code vs. prose tasks). Importantly, because these statistical tests are conducted on a voxel-by-voxel basis (covering tens of thousands of voxels), we apply a *false discovery rate* (FDR) correction to protect against spuriously-significant effects across the brain.

Preprocessing — Removing Noise. The preprocessing step removes noise (such as from motion during the scan). We employed a robust preprocessing pipeline using the Statistical Parametric Mapping 12 Matlab package [115]. First, the functional timeseries were slice-time corrected — this accounts for the fact that *interleaved* slice acquisition during scanning causes slight differences in the relative timing of data collection within a TR (i.e., the 800 ms interval during which whole-brain volumes are sampled). Next, we applied head motion correction and *unwarped* the data using *voxel displacement maps* derived from the fieldmap sequence (see Section 3.4). This step is arguably the most crucial aspect of preprocessing, as head motion is the leading cause of signal artifacts in fMRI data, further interacting with baseline distortion in the magnetic field to geometrically warp voxels. We then segmented and skull-stripped the anatomical images, which were subsequently coregistered to the functional data; both anatomical and functional scans were then spatially-normalized to the Montreal Neurological Institute (MNI152) template [69]. Finally, we constructed a *brain mask* for each participant, which ensures the exclusion of voxels outside of brainspace during statistical analysis.

First-level Analysis — Within One Participant. The first-level analysis focuses on each participant individually. We specified four GLMs for each participant (corresponding to each of the FITB and LR code and prose tasks). The onsets and durations of each trial were defined and convolved with the canonical hemodynamic response function [77] to yield a predicted timeseries of activity (i.e., how we would expect the signal in a voxel to behave if it were sensitive to our task). The data were high-pass filtered (σ = 128s) to remove low-frequency noise, and the model was fit using *robust weighted least squares* (rWLS) [32]. Since these data may be more susceptible to head motion (as a result of typing on the keyboard), we view rWLS as essential for ensuring unbiased parameter estimates: the objective function first obtains an estimate of the noise variance at each scan, and the model is subsequently re-fit after reweighting the data by a factor of 1/variance. Thus, any scans biased by head motion are given less influence in the model, allowing for homogeneous error variance and optimal parameter estimates.

Second-level Analysis — Between Participants. The second-level analysis compares how different participants approached the same task. Prior to second-level GLM, the beta images for each participant were spatially smoothed using a 5 mm^3 full-width at half-maximum (FWHM) Gaussian kernel. These were submitted to an omnibus model (i.e., a factorial analysis of variance) fit using restricted maximum likelihood (ReML). To test for average differences in activity between conditions, we specified several t-contrasts: Code > Prose, FITB Code > FITB Prose, and LR Code > LR Prose. The *contrast* $A > B$ refers to the comparison between task conditions A and B: voxels or features that are more sensitive to A rather than B, or that drive the modeled distinction between A and B. In general, fMRI cannot be used to examine a condition C directly; a subtractive controlled experiment is used instead to compute $A - B$. For example, in our experiments both the FITB and LR tasks feature reading a written prompt, but in general the neural activity associated with reading the prompt "cancels out" when the two are contrasted, and any remaining difference can be attributed to non-identical parts of the experimental condition (i.e., writing code vs. writing prose). The ultimate result of this process is a *statistical parametric map* that displays significant contrast-related activity across the brain, quantified using t-statistics — the *magnitude* of the mean difference between A and B, scaled by model error. Traditionally, brain regions showing significantly more activity in A relative to B are represented with a gradient of 'hot' colors (red to yellow), while regions that are more active during B than A are represented by a gradient of 'cool' colors (blue to green). Such contrast-based analyses are standard for fMRI [39, 52, 103]. All results were FDR-corrected ($q < .05$) and thresholded for a minimum cluster extent of 20 voxels.

5 RESULTS

We analyze our results with respect to four research questions:

RQ1. Do self reports claim code writing is like prose writing?
RQ2. Does the brain treat code writing like prose writing?
RQ3. What low-level features explain code and prose writing?
RQ4. What high-level features explain code and prose writing?

To guide the interpretation of our results, we consider an informal model in which long response coding (the task we studied that is most indicative of coding practice) is made up of the iterative, low-level selection of individual pieces of syntax guided by top-down control. That is, writing a small procedure (the long response task) consists of repeatedly writing the next individual word (the fill-in-the-blank task) while guided by a higher-level goal. Examining the FITB task sheds light on the lower-level basis for code writing, while examining the LR task may illuminate aspects of higher-level "creativity" at the heart of software engineering.

In our fMRI analyses, after filtering incomplete and noisy brain scans, we used data from 24 (8 female, 16 male) of the 30 participants in our experiment. When reporting patterns of neural activity we make use of the standard Brodmann anatomical classification system, which divides the brain into 52 areas (BA 1 through BA 52) [43] based on cytoarchitectural (i.e., cellular-level) similarity. The fMRI results discussed in this section are obtained following the contrast-based analysis methodology described in Section 4.

5.1 RQ1 — Self-Reporting on Code and Prose

We conducted a qualitative analysis of participants' self-reported post survey data. Of our 30 participants, 26 provided their interpretations of similarities between prose and code writing tasks in the post survey. Over a third (38.5%) of these participants reported some similarity between code writing and prose writing. Representative examples include explaining how "filling in the blank

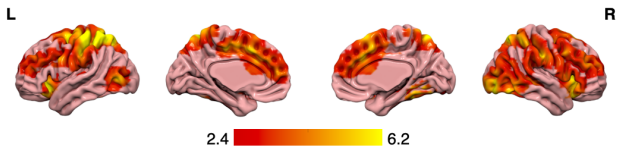


Figure 5: Code writing vs. prose writing contrast. Hotter colors indicate greater t-values (i.e., more activity during coding relative to prose).

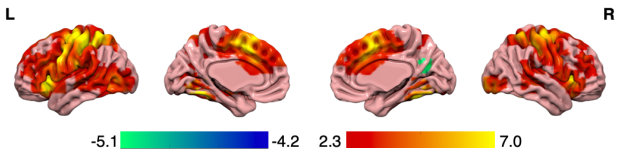


Figure 6: Fill-in-the-blank code vs. fill-in-the-blank prose contrast. Hotter colors indicate more activity during coding relative to prose; cooler colors indicate the reverse.

was like adding variables in code" (we investigate such similarities in Section 5.3) and that both tasks "use logic" (we consider mental representations and problem solving in Section 5.4). Another participant attributed similarity between the two tasks to having "already formed" an idea of the solution that had only to be translated to text (we consider working memory and attentional control in Section 5.2).

As our imaging results will reveal, these subjective reports do not align with measurements of the neural correlates of code and prose writing. Unreliable self-reporting is well-established in both computer science [31, 41, 52, 96] and psychology [66, 90], highlighting the need to augment surveys with more objective metrics.

5.2 RQ2 — Code Writing vs. Prose Writing

We investigate whether there are *general* differences in neural activity between writing code and writing prose. We thus consider all of our code writing tasks (FITB and LR) against all of our prose writing tasks (FITB and LR). This broader Code > Prose contrast, shown in Figure 5, revealed a widely-distributed set of brain regions showing significantly greater activity when writing code. Only significant regions are shown: the colors correspond to the t statistic, which measures the size of the difference relative to the variation in the sample data (t values closer to 0 are not significant after FDR-correction). While care must be taken when comparing such statistics across experiments, as an example baseline we note that the greatest t-value reported in Huang *et al.*'s fMRI study of data structures and spatial ability was $2.0 \leq t \leq 4.2$ [52, Fig. 5]. We view our $2.4 \leq t \leq 6.2$ contrast as a very strong result.

In detail, a particularly large cluster peaked near the left postcentral gyrus and superior parietal lobule (BA 5), extending forward through the primary motor cortex (BA 4) and the premotor/supplementary motor cortex (BA 6). This pattern was also observed in the right hemisphere, albeit yielding somewhat smaller differences in activity (reflected in the smaller t-statistics). However, the right hemisphere did demonstrate more diffuse activity through the lateral prefrontal cortex, including the superior and middle frontal gyri (BA 9–10). The right hemisphere further showed wider clusters of activity in the lateral temporo-occipital and temporoparietal cortex, spanning from the inferior and middle temporal gyri dorsally to the angular gyrus, supramarginal gyrus, and inferior parietal lobule (BA 18–19, 39–40). Finally, we observed comparable patterns of activity in bilateral anterior insula (BA 13) and across the midline of the brain, particularly the medial face of the supplementary motor area (BA 6) and the cingulum (both middle and anterior; BA 24, 32).

> We find a significant ($2.4 \leq t \leq 6.2$) and widely-distributed difference in neural activity between code writing and prose writing in general. The brain does *not* treat code writing and prose writing as similar tasks.

5.3 RQ3 — Code and Prose Foundations

Having established that the brain treats code writing and prose writing differently, we focus attention on our lower-level tasks to explain that difference. We thus consider the contrast FITB Code > FITB Prose, shown in Figure 6. While there was considerable overlap between this contrast and the general Code > Prose analysis (informally, we expect some similarity between writing one word and writing a full sentence), we find that focusing on FITB Code > FITB Prose reveals even stronger ($-5.1 \leq t \leq -4.2$ and $2.3 \leq t \leq 7.0$; conservatively thresholded for multiple comparisons) differences in activity across a number of regions. For example, we observed strong bilateral activity across the entirety of both precentral and postcentral gyri (i.e., the primary motor and somatosensory cortices, respectively; BA 1–4). While these areas are essential for somatomotor function, they are not *cognitive* — that is, activity in these regions does not directly involve 'thought' or 'planning'. These aspects of motor behavior are generally supported by the dorsal premotor cortex and (pre-)supplementary motor area (BA 6, 8), which show significantly greater bilateral activity when performing FITB Code vs. Prose trials. *This suggests that the production of even a single element of code may require more careful, top-down control to effectively plan and produce a context-relevant answer.* This is further supported by significant differences in activity along the frontal eye fields, including the prefrontal eye fields and supplementary eye fields (BA 8–9): these regions are known to help guide the eyes toward relevant stimulus features to generate an appropriate motor plan [49, 116].

We additionally observed significant increases in activity within other regions comprising the so-called 'dorsal attention network' (of which the frontal eye fields are a part). This includes the superior parietal lobule and intraparietal sulcus (BA 7) — structures critical for guiding and maintaining attention in a top-down fashion [23, 61]. Although not part of the dorsal attention network, the bilateral activity found in the anterior insula (BA 13) further supports the notion that FITB Code likely requires more careful

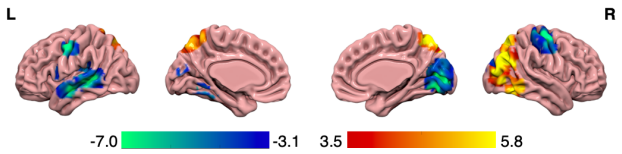


Figure 7: Long response code vs. long response prose contrast. Hotter colors indicate more activity during coding relative to prose; cooler colors indicate the reverse. This represents a strong and exciting result: a significant lateralized difference between prose writing (canonical left hemisphere language areas) and code writing (right hemisphere attention, memory, planning and spatial ability areas).

monitoring of the relevant information needed to provide the appropriate response.

Finally, we note significant differences in activity along posterior temporal/occipital-temporal regions. In general, these appear left dominant, although bilateral activations emerged in the posterior superior temporal gyrus and superior temporal sulcus (BA 21–22). Interestingly, we also observed bilateral activity in the ventral temporal cortex, including the fusiform gyrus (BA 20, 37). While the fusiform gyrus is perhaps best known for its role in face perception, it (along with other areas of the ventral temporal cortex) is also heavily involved in stimulus categorization, particularly for stimuli with which one has developed expertise. This poses the possibility that code — despite being a collection of numbers, letters, and words — is nevertheless treated as a categorically distinct visual stimulus compared to English prose.

At a low level, writing code requires significantly ($-5.1 \leq t \leq -4.2$ and $2.3 \leq t \leq 7.0$) more activity in parts of the brain associated with careful top-down control, planning, and categorization than does writing prose.

5.4 RQ4 — High-Level Coding vs. Prose Writing

Finally, but perhaps most excitingly, we analyze long response code and prose writing tasks. Long response tasks (i.e., writing an entire method) are the most indicative of critical aspects of real-world software engineering. If we consider long response coding to include both the iterative production of single code elements as well as top-down attentional cover of the overarching process, then any difference between this analysis and RQ3 reveals the neurological correlates of that high-level "creativity" in coding.

Figure 7 shows the LR Code > LR Prose trials contrast. This analysis remains strongly significant ($-7.0 \leq t \leq -3.1$ and $3.5 \leq t \leq 5.8$; conservatively thresholded for multiple comparisons) and more precisely pinpoints particular regions. Note how the regions associated with high t-values (hotter colors, more active for code than prose) are largely localized to the right side of the brain. Dually, note how the left hemisphere largely features regions with very low t-values (cooler colors, more active for prose than code). In cognitive neuroscience, such a left vs. right distinction is called *lateralization*. These contrast-based results provide powerful evidence that the *production* of code vs. prose relies on highly distinct cognitive substrates.

Prose production was strongly associated with left temporal regions classically associated with natural language (which is almost entirely left-lateralized in right-handed individuals). Namely, we saw increased recruitment of the middle temporal gyrus (MTG) and superior temporal gyrus (STG) (BA 21–22). The left MTG has previously been shown to activate when accessing semantic aspects of language and is thought to support a lexicon of words [59, 82]. The STG extends into Wernicke's area, which is notably the primary center of language comprehension [21]. Although it generally appears most active during comprehension of *spoken* language, the act of writing often involves a sort of internal narration that may similarly recruit these regions. This is further supported by increased activation of the calcarine (visual) cortex, particularly the lingual gyrus along the right medial wall (BA 17–18). The lingual gyrus, while not playing a role in higher-order language processes *per se*, is often associated with the recognition of letters and words, perhaps contributing to their semantic understanding [72]. We also observed a small cluster of activity in the inferior frontal gyrus (BA 44) — part of Broca's area, which underlies the production of language (although, again, is more commonly linked to speech) [37].

Code production, by contrast, was largely right-lateralized. The exception to this observation was a bilateral activation of the superior parietal lobule, extending dorsolaterally into the precuneus along the midline (BA 7). The superior parietal lobule (see Section 5.3) is involved in top-down control processes related to attention and memory; the precuneus is associated with processes such as mental imagery [118]. Similarly, we observed right temporal and temporoparietal activations along a number of regions supporting visual association (tying visual information together) and other forms of mental imagery, including spatial cognition (BA 19, 39). The angular gyrus, in particular, may support various aspects of spatial and mathematical reasoning, including the manipulation of mental representations and other aspects of problem-solving [45]. Importantly, it is thought to act as a bridge between perception, recognition, and action, suggesting that code synthesis may require a more complex interplay of understanding a problem and formulating a comprehensive plan to solve it [100]. This swath of activity extended ventrally into regions of the inferior occipital-temporal cortex, which partially overlap with clusters identified by Huang *et al.* as being more active during difficult data structure manipulations (relative to difficult mental rotation tasks) [52]. Together, these findings suggest that code production is perhaps more 'spatial' in nature, requiring the formulation of a mental map that guides problem-solving.

Very informally, finding activity in the (expected, standard) language areas for prose writing gives us high confidence that we designed and carried out our controlled experiment correctly in general. However, that high confidence makes the observation that long-form code writing does not heavily recruit these areas (instead using parts of the brain associated with planning and spatial ability) all the more startling. Part of the motivation for Siegmund *et al.*'s pioneering first use of fMRI in software engineering [103] was to provide direct evidence, one way or another, regarding claims such as Dijkstra's that "exceptionally good mastery of one's native

tongue is the most vital asset of a competent programmer" [33]. While that may be true for code reading (e.g., comprehension [103] and reviewing [39]), our results suggest that it is *not* true for code writing at a neural level.

> High-level long response coding is significantly different ($-7.0 \leq t \leq -3.1$ and $3.5 \leq t \leq 5.8$) from prose writing. Prose writing involves areas of the brain canonically associated with language. Coding involves a different set of right-lateralized regions associated with attention, memory, planning, and spatial ability. This provides the first evidence of significant neural differences between *prose writing* (which is neurally similar to natural language) and *code writing* (which, we find, is *not*).

5.5 Summary of Results

At a high level, an analysis of all code writing tasks against all prose writing tasks showed that the two operate via distinct neural mechanisms. We analyzed these differences at a more granular level by considering imaging data from tasks of the same type (i.e., FITB, LR). The FITB Code > FITB Prose contrast established the low-level cognitive features distinct to code writing: brain regions associated with top-down control, planning, and categorization. Subsequent analyses of LR tasks revealed a clear lateralized distinction between code writing and prose writing. Largely, we found that code writing involves right hemisphere brain regions involved in spatial ability and planning while prose writing involves the canonical left hemisphere regions associated with language production. In addition to supporting previous medical imaging studies of prose writing and software engineering tasks, these findings introduce a new and alternative relationship between code and prose in which reading and writing are *not* similar (cf. [33, 39, 103]).

6 THREATS TO VALIDITY

Our choice of writing tasks presents a first potential threat to the validity of our experiment. While various forms of code writing exist in software engineering contexts (e.g., testing, debugging), we restricted our task set to *prompted* code writing tasks. We mitigate the threat that this limited benchmark poses via our robust experimental design, whereby participants complete different types of writing tasks (i.e., FITB, LR). We further address this concern by including a variety of fundamental programming concepts (e.g., both control- and data-flow operations) in our selected coding tasks indicative of many real-world coding tasks. Nevertheless, our results may not generalize to all in-the-wild programming; we leave a more thorough investigation to future research.

Secondly, the design of our tasks may have impacted our ability to measure brain activity strictly associated with code and prose writing. For example, our stimuli included written instructions that participants read before typing their responses. This construction introduces the possibility that we measure brain activity beyond strictly writing responses. We designed our contrast-based experiment to mitigate this threat. As fMRI analyses are subtractive (described above in Section 5), the effects of reading the prompt cancel out, leaving only the differences between prose writing and code writing. However, we note that differences exist in the prompt text contained in FITB stimuli (i.e., Prose FITB and Code FITB tasks require the participant to read prose and code, respectively, see Figures 2a and 2c). Overall, we maintain that FITB tasks measure the process of low-level selection of individual code elements, a distinct activity to pure comprehension.

Lastly, our results may be limited by our participant cohort. For this experiment, we recruited undergraduate and graduate students with an average of 5.2 semesters of programming experience. Thus, our results may not extend to programmers with different backgrounds or expertise. Indeed, previous fMRI studies have investigated the role of expertise and demographics in detail (e.g., [18, 39]). We claim no significant findings regarding individual differences and report results for our participants as a whole.

7 RELATED WORK

In this section we place our contributions in context, comparing our technique and results to other approaches.

7.1 Medical Imaging and Software Engineering

As of 2019, the application of medical imaging to understand the cognitive processes associated with computer science is still in its infancy. In a 2014 study, Siegmund *et al.* were the first to use fMRI in the context of software engineering, identifying five brain regions associated with code comprehension [103]. Peitek *et al.* conducted follow up studies of program comprehension, using fMRI to study programmers' cognitive load [83] and the neural efficiency of top-down and bottom-up methods [85]. Researchers have also adopted alternative medical imaging modalities to study code comprehension, including functional near-infrared spectroscopy (fNIRS) [36, 75], electroencephalography (EEG) [25, 63], and electromyography (EMG) [79]; we selected fMRI in part because recent research suggests it is necessary to find neural correlates of some subtle programming activities [52].

Of the previous medical imaging studies of software engineering, we consider Floyd *et al.*'s investigation of code comprehension, code review, and prose review to be the closest to our work. Their study contrasted the cognitive processes of code and prose review tasks, finding the neural representations of natural and programming languages to be distinct, but that those differences are modulated by expertise [39]. However, no previous study, including Floyd *et al.*'s work, has used fMRI to study code writing. Floyd *et al.*'s previous use of natural language as a baseline, combined with Huang *et al.*'s findings on the importance of fMRI's spatial resolution to study software engineering tasks [52], serves as supporting evidence for our experimental design.

7.2 Code Writing and Cognition

Developing methods to make code writing more productive, accurate, and accessible has long been a software engineering goal. Such methods are often motivated by an understanding of programming psychology, such as block-style languages designed for younger ages (e.g., Scratch [93]) and IDEs intended to increase productivity [108].

Researchers from as early as the 1950s have sought to understand the psychology of writing code. Early efforts focused on cognitive load [14], psycholinguistic theory [105], and expertise [46, 117], among other topics. In 1977, Brooks proposed an entire theory of

programming behavior oriented toward explaining transcriptions of participants asked to talk aloud while performing programming tasks [13]. More recent work includes studies on how experts and novices classify algorithms [78, 114] and programmers' use of the Web when writing code [12]. In contrast to our work, all previous studies used observational techniques (i.e., they were unable to take advantage of medical imaging).

7.3 Prose Writing and Cognition

Like code writing, early research dedicated to the cognitive processes of prose writing was conducted without medical imaging. Similar to Brooks' theoretical framework of the coding process, Hayes and Flowers proposed a theory of the cognitive processes of writing in 1981 [38]. Research in the field continued throughout the 80's and 90's, focusing on more nuanced aspects of prose writing cognition, including second-language proficiency [26, 64] and studies on gaining writing expertise (e.g., [6]).

Unlike code writing, researchers have since leveraged medical imaging to establish objective models of the prose writing process and have used such an understanding to improve prose writing as a whole. Menon and Desmond were among the first to use fMRI to understand prose writing. Their study, in which participants wrote by dictation in an fMRI machine, found activation in only the left hemisphere, particularly the left superior parietal lobe [73]. Our study similarly found activation in left temporal region for prose, but also found the right temporal region to be associated with code writing. Shah *et al.* later used fMRI to study the neural correlates of creative writing with an experiment that separated the "brainstorming" phase of prose writing from "the act of writing a new and creative continuation of a given literary text" [101]. We consider this more analogous to the planning and implementation stages of the software engineering lifecycle, and note that our experiments cover the act of code writing but not explicit planning.

There has also been particular interest in studying the specialization of writing-specific brain regions. Sugihara *et al.* studied the brain's writing center during both left- and right-handed writing tasks [111], identifying regions crucial to the core process of writing. Planton *et al.* later identified brain regions that are consistently involved in prose writing tasks, as well as differences in brain activation across writing, drawing, and oral spelling [89]. Similarly, Purcell *et al.* studied the neural basis of spelling and its relation that of reading: their study used a QWERTY keyboard to study *prose* writing with fMRI [92]. However, their experimental design was restricted to typing single words by dictation and without the participants having any live feedback while typing.

Historically, the development of a fundamental, neurological understanding of other activities, such as prose writing, has proved useful as a guide in pedagogy and research. Berninger and Winn credit advanced brain-imaging technologies as the primary development near the end of the 20th century that reformed prose writing research and education [10]. Examples of brain imaging contributing to pedagogy include the use of verbal and non-verbal cues and strategies to improve learning [51, 58], as well as teaching such cues and strategies to overcome inefficiencies in temporally-constrained verbal working memory [112]. In comparison, researchers lack a corresponding fundamental understanding of code writing that might illuminate new ways to improve code writing skills. Our work is partially motivated by the belief that such a foundational understanding could guide more focused training and teaching strategies for code writing.

7.4 Code Writing and Prose Writing

Despite apparent differences (e.g., syntax) between prose writing and code writing, previous research has found connections between the two activities. Such research largely focuses on the use of code writing or prose writing as a tool in the instruction of the other.

Kelleher and Pausch conducted a seminal study to investigate this pedagogical approach in 2007, using the Storytelling Alice programming environment to inspire middle school girls' interest in programming [56]. Later, Burke extended this work to a formal classroom setting with the language Scratch, highlighting the similarities in product, process, and perception between the code and prose writing [16].

In the other direction, in a 2010 study, Burke and Kafai used computer program writing to help children develop their storytelling and creative writing abilities [17]. The study highlighted the shared characteristics of sequence, structure, and clarity of expression between the two activities, and emphasized the utility of programming as a means of reflecting on the storytelling process.

These studies form a call to arms for a more thorough investigation into the relationship between code and prose production, but no previous paper has examined that fundamental writing relationship using medical imaging. The stark differences that we present between the neurological representations of code and prose writing, in light of the similarities reported by these previous studies, highlight the need for such quantitative research.

8 CONCLUSION

Over the decades, researchers from Dijkstra and Pausch to Pea and Kurland, among many others, have made observational investigations into, theories of, and calls to arms regarding the psychological aspects of programming. Understanding cognition has helped improve prose reading, prose writing, and code reading — but code writing has lacked neurologically-grounded, indicative research. Indeed, since the first fMRI study of software engineering five years ago, there have been fewer than a dozen fMRI experiments reported at major SE conferences [22, 34, 36, 39, 52, 53, 75, 84, 84, 103, 104].

We present the first fMRI study of code writing. We employ a controlled, contrast-based experiment in which code writing, prose writing, fill-in-the-blank and long response tasks are presented to participants, who make use of a special fMRI-safe keyboard to type their responses in a realistic live editing setting.

We report three primary results. First, there is a significant and widely-distributed difference in neural activity between code writing and prose writing in general: the brain does *not* treat code writing and prose writing as similar tasks. Second, at a low level (i.e., producing a single word or code element), writing code requires significantly more activity in brain areas associated with careful, top-down control, planning and categorization: despite superficial similarity, code appears to be a categorically distinct visual stimulus compared to prose. Third, and most excitingly, high-level long response coding — the studied task perhaps most indicative

of real-world programming — is significantly different from prose writing. While prose writing involves left-brain regions canonically associated with language, we find a sharp lateralized distinction: code writing does *not* significantly recruit those regions compared to prose writing, instead showing activation in right-brain areas associated with attention, memory, planning and spatial ability. While previous studies have found that code and prose *reading* may be similar at an observational or neurological level, **we present the first evidence suggesting that code and prose *writing* are quite dissimilar at the neurological level**. This unexpected result — that the production of code and prose rely on highly distinct cognitive substrates — though quite preliminary, paves the way for future investigations analogous to those based on medical imaging for prose writing. In addition to developing a foundational understanding of code-writing, this empirical distinction may be leveraged to develop tools and pedagogies (e.g., transfer training), subsequently affecting large scale workforce retraining and educational reform. Moreover, neurological evidence that code and prose writing are not as intertwined as conventionally thought may encourage more diverse participation in computer science.

ACKNOWLEDGEMENTS

We thank Kevin Angstadt, Ian Bertram, Madeline Endres, Michael Flanagan, Nicholas McKay, and Zohreh Sharafi for their help managing participants and with feedback and revisions to early drafts.

REFERENCES

[1] B. P. Acevedo, E. N. Aron, A. Aron, M.-D. Sangster, N. Collins, and L. L. Brown. The highly sensitive brain: an fMRI study of sensory processing sensitivity and response to others' emotions. *Brain and behavior*, 4(4):580–594, 2014.

[2] D. Arnow and O. Barshay. On-line programming examinations using web to teach. In *Conference on Innovation and Technology in Computer Science Education*, pages 21–24, 1999.

[3] D. Arnow and G. Weiss. Turing's craft. In *https://www.turingscraft.com/*, 2019.

[4] M. Atherton, J. Zhuang, W. M. Bart, X. Hu, and S. He. A functional MRI study of high-level cognition. I. The game of chess. *Cognitive Brain Research*, 16(1):26–31, 2003.

[5] D. S. Bassett, N. F. Wymbs, M. A. Porter, P. J. Mucha, J. M. Carlson, and S. T. Grafton. Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 108(18):7641–7646, 2011.

[6] A. Beaufort. Learning the trade: A social apprenticeship model for gaining writing expertise. *Written communication*, 17(2):185–223, 2000.

[7] M. Behroozi, A. Lui, I. Moore, D. Ford, and C. Parnin. Dazed: measuring the cognitive load of solving technical interview problems at the whiteboard. In *International Conference on Software Engineering: New Ideas and Emerging Results (ICSE NIER)*, pages 93–96, 2018.

[8] C. M. Bennett, M. Miller, and G. L. Wolford. Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for proper multiple comparisons correction. *NeuroImage*, 47, July 2009.

[9] V. W. Berninger, T. L. Richards, P. S. Stock, R. D. Abbott, P. A. Trivedi, L. E. Altemeier, and J. R. Hayes. fMRI activation related to nature of ideas generated and differences between good and poor writers during idea generation. In *BJEP Monograph Series II, Number 6-Teaching and Learning Writing*, volume 77, pages 77–93. British Psychological Society, 2009.

[10] V. W. Berninger and W. Winn. Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. *Handbook of writing research*, pages 96–114, 2006.

[11] J. L. Booth and R. S. Siegler. Numerical magnitude representations influence arithmetic learning. *Child Development*, 79(4):1016–1031, 2008.

[12] J. Brandt, P. J. Guo, J. Lewenstein, M. Dontcheva, and S. R. Klemmer. Two studies of opportunistic programming: interleaving web foraging, learning, and writing code. In *Conference on Human Factors in Computing Systems*, pages 1589–1598, 2009.

[13] R. Brooks. Towards a theory of the cognitive processes in computer programming. *International Journal of Man-Machine Studies*, 9(6):737–751, 1977.

[14] J. S. Bruner, J. J. Goodnow, and G. A. Austin. *A study of thinking*. New York: John Wiley & Sons, Inc, 1956.

[15] J. A. Bugos, W. M. Perlstein, C. S. McCrae, T. S. Brophy, and P. H. Bedenbaugh. Individualized piano instruction enhances executive functioning and working memory in older adults. *Aging and Mental Health*, 11(4):464–471, 2007.

[16] Q. Burke. The markings of a new pencil: Introducing programming-as-writing in the middle school classroom. *Journal of Media Literacy Education*, 4(2):121–135, 2012.

[17] Q. Burke and Y. B. Kafai. Programming & storytelling: opportunities for learning about coding & composition. In *Proceedings of the 9th International Conference on Interaction Design and Children*, pages 348–351. ACM, 2010.

[18] T. Busjahn, R. Bednarik, A. Begel, M. Crosby, J. H. Paterson, C. Schulte, B. Sharif, and S. Tamm. Eye movements in code reading: Relaxing the linear order. In *International Conference on Program Comprehension*, pages 255–265, 2015.

[19] R. Cabeza, S. M. Daselaar, F. Dolcos, S. E. Prince, M. Budde, and L. Nyberg. Task-independent and task-specific age effects on brain activity during working memory, visual attention and episodic retrieval. *Cerebral Cortex*, 14(4):364–375, 2004.

[20] S. Caminiti. AT&T's $1 billion gambit: Retraining nearly half its workforce for jobs of the future. In *https://www.cnbc.com/2018/03/13/atts-1-billion-gambit-retraining-nearly-half-its-workforce.html*, Mar 2018.

[21] R. Campbell. The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):1001–1010, 2007.

[22] J. Castelhano, I. C. Duarte, C. Ferreira, J. Duraes, H. Madeira, and M. Castelo-Branco. The Role of the Insula in Intuitive Expert Bug Detection in Computer Code: An fMRI Study. *Brain Imaging and Behavior*, May 2018.

[23] M. Corbetta, G. L. Shulman, F. M. Miezin, and S. E. Petersen. Superior parietal cortex activation during spatial attention shifts and visual feature conjunction. *Science*, 270(5237):802–805, 1995.

[24] K. Costello and G. Omale. Gartner says global it spending to reach $3.8 trillion in 2019. In *https://www.gartner.com/en/newsroom/press-releases/2019-01-28-gartner-says-global-it-spending-to-reach--3-8-trillio*, Jan 2019.

[25] I. Crk, T. Kluthe, and A. Stefik. Understanding programming expertise: an empirical study of phasic brain wave changes. *Transactions on Computer-Human Interaction*, 23(1):2, 2016.

[26] A. Cumming. Writing expertise and second-language proficiency. *Language learning*, 39(1):81–135, 1989.

[27] C. Cutter. Amazon to retrain a third of its U.S. workforce. In *https://www.wsj.com/articles/amazon-to-retrain-a-third-of-its-u-s-workforce-11562841120*, Jul 2019.

[28] R. G. Cuya, G. Nivera, and E. C. Fortes. The use of non-math analogies in teaching mathematics. *The Normal Lights*, 11(1), 2017.

[29] E. Dahlin, A. S. Neely, A. Larsson, L. Bäckman, and L. Nyberg. Transfer of learning after updating training mediated by the striatum. *Science*, 320(5882):1510–1512, 2008.

[30] B. De Smedt, D. Ansari, R. H. Grabner, M. M. Hannula, M. Schneider, and L. Verschaffel. Cognitive neuroscience meets mathematics education. *Educational Research Review*, 5(1):97–105, 2010.

[31] N. Dell, V. Vaidyanathan, I. Medhi, E. Cutrell, and W. Thies. Yours is better!: participant response bias in HCI. In *Human Factors in Computing Systems*, pages 1321–1330. ACM, 2012.

[32] J. Diedrichsen and R. Shadmehr. Detecting and adjusting for artifacts in fMRI time series data. *NeuroImage*, 27(3):624–634, 2005.

[33] E. Dijkstra. How do we tell truths that might hurt? In *Selected Writings on Computing: A Personal Perspective*, pages 129–131. Springer, 1982.

[34] J. Duraes, H. Madeira, J. Castelhano, C. Duarte, and M. C. Branco. WAP: Understanding the Brain at Software Debugging. In *International Symposium on Software Reliability Engineering*, pages 87–92, 2016.

[35] R. English and G. Edwards. Programming as a writing activity. *Computing Teacher*, 11(6):46–47, 1984.

[36] S. Fakhoury, Y. Ma, V. Arnaoudova, and O. Adesope. The effect of poor source code lexicon and readability on developers' cognitive load. In *International Conference on Program Comprehension*, pages 286–296, 2018.

[37] A. Flinker, A. Korzeniewska, A. Y. Shestyuk, P. J. Franaszczuk, N. F. Dronkers, R. T. Knight, and N. E. Crone. Redefining the role of Broca's area in speech. *Proceedings of the National Academy of Sciences*, 112(9):2871–2875, 2015.

[38] L. Flower and J. R. Hayes. A cognitive process theory of writing. *College Composition and Communication*, 32(4):365–387, 1981.

[39] B. Floyd, T. Santander, and W. Weimer. Decoding the representation of code in the brain: An fMRI study of code review and expertise. In *International Conference on Software Engineering*, pages 175–186, 2017.

[40] M. Fowler. *Refactoring: improving the design of existing code*. Addison-Wesley Professional, 2018.

[41] Z. P. Fry, B. Landau, and W. Weimer. A human study of patch maintainability. In *International Symposium on Software Testing and Analysis*, pages 177–187, 2012.

[42] W. D. Gaillard, B. C. Sachs, J. R. Whitnah, Z. Ahmad, L. M. Balsamo, J. R. Petrella, S. H. Braniecki, C. M. McKinney, K. Hunter, B. Xu, et al. Developmental aspects of language processing: fMRI of verbal fluency in children and adults. *Human*

Brain Mapping, 18(3):176–185, 2003.
[43] L. J. Garey. *Brodmann's "Localisation in the Cerebral Cortex"*. World Scientific, 2006.
[44] G. H. Glover. Overview of functional magnetic resonance imaging. *Neurosurgery Clinics*, 22(2):133–139, 2011.
[45] R. H. Grabner, A. Ischebeck, G. Reishofer, K. Koschutnig, M. Delazer, F. Ebner, and C. Neuper. Fact learning in complex arithmetic and figural-spatial tasks: The role of the angular gyrus and its relation to mathematical competence. *Human brain mapping*, 30(9):2936–2952, 2009.
[46] E. E. Grant and H. Sackman. An exploratory investigation of programmer performance under on-line and off-line conditions. *IEEE Transactions on Human Factors in Electronics*, (1):33–48, 1967.
[47] L. W. Gregg and E. R. Steinberg. *Cognitive processes in writing*. Routledge, 2016.
[48] J. Halberda, M. M. Mazzocco, and L. Feigenson. Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455(7213):665, 2008.
[49] H. R. Heekeren, S. Marrett, P. A. Bandettini, and L. G. Ungerleider. A general mechanism for perceptual decision-making in the human brain. *Nature*, 431(7010):859, 2004.
[50] Y. Higashiyama, K. Takeda, Y. Someya, Y. Kuroiwa, and F. Tanaka. The neural basis of typewriting: A functional MRI study. *PLOS ONE*, 10(7):1–20, 07 2015.
[51] S. R. Hooper, L.-J. C. Costa, M. McBee, K. L. Anderson, D. C. Yerby, A. Childress, and S. B. Knuth. A written language intervention for at-risk second grade students: a randomized controlled trial of the process assessment of the learner lesson plans in a tier 2 response-to-intervention (RtI) model. *Annals of Dyslexia*, 63(1):44–64, 2013.
[52] Y. Huang, X. Liu, R. Krueger, T. Santander, X. Hu, K. Leach, and W. Weimer. Distilling neural representations of data structure manipulation using fMRI and fNIRS. In *International Conference on Software Engineering*, pages 396–407, 2019.
[53] Y. Ikutani and H. Uwano. Brain activity measurement during program comprehension with NIRS. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pages 1–6. IEEE, 2014.
[54] G. A. James, G. He, and Y. Liu. A full-size MRI-compatible keyboard response system. *Neuroimage*, 25(1):328–31, Mar. 2005.
[55] L. Kaufmann, S. E. Vogel, G. Wood, C. Kremser, M. Schocke, L.-B. Zimmerhackl, and J. W. Koten. A developmental fMRI study of nonsymbolic numerical and spatial processing. *Cortex*, 44(4):376–385, 2008.
[56] C. Kelleher and R. Pausch. Using storytelling to motivate programming. *Communications of the ACM*, 50(7):58–64, 2007.
[57] R. T. Kellogg. Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, 1(1):1–26, 2008.
[58] R. T. Kellogg and A. P. Whiteford. Training advanced writing skills: The case for deliberate practice. *Educational Psychologist*, 44(4):250–266, 2009.
[59] D. Kemmerer. Word classes in the brain: Implications of linguistic typology for cognitive neuroscience. *Cortex*, 58:27–51, 2014.
[60] D. E. Knuth. *The art of computer programming*, volume 3. Pearson Education, 1997.
[61] M. Koenigs, A. K. Barbey, B. R. Postle, and J. Grafman. Superior parietal cortex is critical for the manipulation of information in working memory. *Journal of Neuroscience*, 29(47):14980–14986, 2009.
[62] K. Landerl and C. Kölle. Typical and atypical development of basic numerical skills in elementary school. *Journal of Experimental Child Psychology*, 103(4):546–565, 2009.
[63] S. Lee, A. Matteson, D. Hooshyar, S. Kim, J. Jung, G. Nam, and H. Lim. Comparing programming language comprehension between novice and expert programmers using EEG analysis. In *International Conference on Bioinformatics and Bioengineering*, pages 350–355, Oct 2016.
[64] I. Leki. Building expertise through sequenced writing assignments. *Teachers of English to Speakers of Other Languages Journal*, 1(2):19–23, 1992.
[65] C. J. Limb and A. R. Braun. Neural substrates of spontaneous musical performance: An fMRI study of jazz improvisation. *PLoS one*, 3(2):e1679, 2008.
[66] P. Mabe and S. West. Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology*, 67(3):280–296, 6 1982.
[67] E. A. Maguire, K. Woollett, and H. J. Spiers. London taxi drivers and bus drivers: A structural MRI and neuropsychological analysis. *Hippocampus*, 16(12):1091–1101, 2006.
[68] S. Maguire. *Writing solid code*. Greyden Press, LLC, 2013.
[69] J. B. A. Maintz and M. A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998.
[70] R. C. Martin. *Clean code: a handbook of agile software craftsmanship*. Pearson Education, 2009.
[71] B. McCandliss, I. L. Beck, R. Sandak, and C. Perfetti. Focusing attention on decoding for children with poor reading skills: Design and preliminary tests of the word building intervention. *Scientific Studies of Reading*, 7(1):75–104, 2003.
[72] A. Mechelli, G. W. Humphreys, K. Mayall, A. Olson, and C. J. Price. Differential effects of word length and visual contrast in the fusiform and lingual gyri during. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267(1455):1909–1913, 2000.
[73] V. Menon and J. Desmond. Left superior parietal cortex involvement in writing: integrating fMRI with lesion evidence. *Cognitive Brain Research*, 12(2):337–340, 2001.
[74] M. P. Milham, K. I. Erickson, M. T. Banich, A. F. Kramer, A. Webb, T. Wszalek, and N. J. Cohen. Attentional control in the aging brain: Insights from an fMRI study of the Stroop task. *Brain and Cognition*, 49(3):277–296, 2002.
[75] T. Nakagawa, Y. Kamei, H. Uwano, A. Monden, K. Matsumoto, and D. M. German. Quantifying programmers' mental workload during program comprehension based on cerebral blood flow measurement: A controlled experiment. In *Companion Proceedings of the 36th International Conference on Software Engineering*, pages 448–451, 2014.
[76] T. Nakao, A. Nakagawa, T. Yoshiura, E. Nakatani, M. Nabeyama, C. Yoshizato, A. Kudoh, K. Tada, K. Yoshioka, M. Kawamoto, O. Togao, and S. Kanba. Brain activation of patients with obsessive-compulsive disorder during neuropsychological and symptom provocation tasks before and after symptom improvement: a functional magnetic resonance imaging study. *Biological Psychiatry*, 57(8):901–910, 2005.
[77] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.
[78] T. Ormerod. Human cognition and programming. In *Psychology of Programming*, pages 63–82. Elsevier, 1990.
[79] C. Parnin. Subvocalization-toward hearing the inner thoughts of developers. In *International Conference on Program Comprehension*, pages 197–200. IEEE, 2011.
[80] R. D. Pea and D. M. Kurland. On the cognitive prerequisites of learning computer programming. Technical Report 18, Center for Children and Technology, 1983.
[81] R. D. Pea and D. M. Kurland. On the cognitive effects of learning computer programming. *New Ideas in Psychology*, 2(2):137–168, 1984.
[82] M. V. Peelen, D. Romagno, and A. Caramazza. Independent representations of verbs and actions in left lateral temporal cortex. *Journal of cognitive neuroscience*, 24(10):2096–2107, 2012.
[83] N. Peitek, J. Siegmund, C. Parnin, S. Apel, and A. Brechmann. Beyond gaze: preliminary analysis of pupil dilation and blink rates in an fMRI study of program comprehension. In *Proceedings of the Workshop on Eye Movements in Programming*, page 4. ACM, 2018.
[84] N. Peitek, J. Siegmund, C. Parnin, S. Apel, J. Hofmeister, and A. Brechmann. Simultaneous Measurement of Program Comprehension with fMRI and Eye Tracking: A Case Study. In *Symposium on Empirical Software Engineering and Measurement*, 2018.
[85] N. Peitek, J. Siegmund, C. Parnin, S. Apel, J. Hofmeister, C. Kästner, A. Begel, A. Bethmann, and A. Brechmann. Neural efficiency of top-down program comprehension. *Software Engineering und Software Management 2018*, 2018.
[86] M. L. Pelchat, A. Johnson, R. Chan, J. Valdez, and J. D. Ragland. Images of desire: food-craving activation during fMRI. *Neuroimage*, 23(4):1486–1493, 2004.
[87] S. J. Pickering and P. Howard-Jones. Educators' views on the role of neuroscience in education: Findings from a study of UK and international perspectives. *Mind, Brain, and Education*, 1(3):109–113, 2007.
[88] S. Planton, M. Jucla, F.-E. Roux, and J.-F. Démonet. The "handwriting brain": a meta-analysis of neuroimaging studies of motor versus orthographic processes. *Cortex*, 49(10):2772–2787, 2013.
[89] S. Planton, M. Longcamp, P. Péran, J.-F. Demonet, and M. Jucla. How specialized are writing-specific brain regions? an fMRI study of writing, drawing and oral spelling. *Cortex*, 88:66–80, 2017.
[90] P. M. Podsakoff and D. W. Organ. Self-reports in organizational research: Problems and prospects. *Journal of Management*, 12(4):531–544, 1986.
[91] Psychology Software Tools, Inc. E-Prime. https://pstnet.com/products/e-prime/.
[92] J. J. Purcell, E. M. Napoliello, and G. F. Eden. A combined fMRI study of typed spelling and reading. *NeuroImage*, 55(2):750 – 762, 2011.
[93] M. Resnick, J. Maloney, A. Monroy-Hernández, N. Rusk, E. Eastmond, K. Brennan, A. Millner, E. Rosenbaum, J. Silver, B. Silverman, and Y. Kafai. Scratch: Programming for all. *Commun. ACM*, 52(11):60–67, Nov. 2009.
[94] T. L. Richards, V. W. Berninger, and M. Fayol. fMRI activation differences between 11-year-old good and poor spellers' access in working memory to temporary and long-term orthographic representations. *Journal of Neurolinguistics*, 22(4):327–353, 2009.
[95] T. L. Richards, V. W. Berninger, P. Stock, L. Altemeier, P. Trivedi, and K. R. Maravilla. Differences between good and poor child writers on fMRI contrasts for writing newly taught and highly practiced letter forms. *Reading and Writing*, 24(5):493–516, 2011.
[96] S. Riley. Password security: What users know and what they actually do. *Usability News*, 8(1):2833–2836, 2006.
[97] M. Ritchey, A. P. Yonelinas, and C. Ranganath. Functional connectivity relationships predict similarities in task activation and pattern information during associative memory encoding. *J. Cognitive Neuroscience*, 26(5):1085–1099, May 2014.
[98] J. S. Ross, J. Tkach, P. M. Ruggieri, M. Lieber, and E. Lapresto. The mind's eye: Functional MR imaging evaluation of golf motor imagery. *American Journal of Neuroradiology*, 24(6):1036–1044, 2003.

[99] V. Sarina and I. K. Namukasa. Nonmath analogies in teaching mathematics. *Procedia-Social and Behavioral Sciences*, 2(2):5738–5743, 2010.
[100] M. L. Seghier. The angular gyrus: multiple functions and multiple subdivisions. *The Neuroscientist*, 19(1):43–61, 2013.
[101] C. Shah, K. Erhard, H.-J. Ortheil, E. Kaza, C. Kessler, and M. Lotze. Neural correlates of creative writing: an fMRI study. *Human Brain Mapping*, 34(5):1088–1101, 2013.
[102] Y. I. Sheline, D. M. Barch, J. M. Donnelly, J. M. Ollinger, A. Z. Snyder, and M. A. Mintun. Increased amygdala response to masked emotional faces in depressed subjects resolves with antidepressant treatment: an fmri study. *Biological Psychiatry*, 50(9):651–658, 2001.
[103] J. Siegmund, C. Kästner, S. Apel, C. Parnin, A. Bethmann, T. Leich, G. Saake, and A. Brechmann. Understanding understanding source code with functional magnetic resonance imaging. In *International Conference on Software Engineering*, pages 378–389, 2014.
[104] J. Siegmund, N. Peitek, C. Parnin, S. Apel, J. Hofmeister, C. Kästner, A. Begel, A. Bethmann, and A. Brechmann. Measuring Neural Efficiency of Program Comprehension. In *Foundations of Software Engineering*, pages 140–150, 2017.
[105] M. Sime, T. Green, and D. Guest. Psychological evaluation of two conditional constructions used in computer languages. *International Journal of Man-Machine Studies*, 5(1):105–113, 1973.
[106] P. G. Simos, J. M. Fletcher, E. Bergman, J. Breier, B. Foorman, E. Castillo, R. Davis, M. Fitzgerald, and A. Papanicolaou. Dyslexia-specific brain activation profile becomes normal following successful remedial training. *Neurology*, 58(8):1203–1213, 2002.
[107] N. Singer. The hard part of computer science? getting into class. In *https://www.nytimes.com/2019/01/24/technology/computer-science-courses-college.html*, Jan 2019.
[108] D. C. Smith. Pygmalion: a creative programming environment. Technical Report STAN-CS-75-499, Stanford University, 1975.
[109] M. Snejbjerg Jensen, O. Heggli, P. Mota, and P. Vuust. A low-cost MRI compatible keyboard. In *International Conference on New Interfaces for Musical Expression*, 2017.
[110] C. J. Stoodley, E. M. Valera, and J. D. Schmahmann. Functional topography of the cerebellum for motor and cognitive tasks: an fMRI study. *Neuroimage*, 59(2):1560–1570, 2012.
[111] G. Sugihara, T. Kaminaga, and M. Sugishita. Interindividual uniformity and variety of the "writing center": a functional MRI study. *Neuroimage*, 32(4):1837–1849, 2006.
[112] C. Titz and J. Karbach. Working memory and executive functions: effects of training on academic achievement. *Psychological Research*, 78(6):852–868, 2014.
[113] S. J. Van Rooij, E. Geuze, M. Kennis, A. R. Rademaker, and M. Vink. Neural correlates of inhibition and contextual cue processing related to treatment response in ptsd. *Neuropsychopharmacology*, 40(3):667, 2015.
[114] M. Weiser and J. Shertz. Programming problem representation in novice and expert programmers. *International Journal of Man-Machine Studies*, 19(4):391–398, 1983.
[115] Wellcome Trust Centre for Neuroimaging. Statistical parametric mapping. In *http://www.fil.ion.ucl.ac.uk/spm/*, Aug. 2016.
[116] S.-n. Yang, H. Hwang, J. Ford, and S. Heinen. Supplementary eye field activity reflects a decision rule governing smooth pursuit but not the decision. *Journal of Neurophysiology*, 103(5):2458–2469, 2010.
[117] E. A. Youngs. Human errors in programming. *International Journal of Man-Machine Studies*, 6(3):361–376, 1974.
[118] S. Zhang and R. L. Chiang-shan. Functional connectivity mapping of the human precuneus by resting state fmri. *Neuroimage*, 59(4):3548–3562, 2012.