Title

Reproducibility of brain-cognition relationships using three cortical surface-based protocols: An exhaustive analysis based on cortical thickness

Journal

Human Brain Mapping, 36(8)

ISSN

1065-9471

Authors

Martínez, Kenia
Madsen, Sarah K
Joshi, Anand A
et al.

Publication Date

2015-08-01

DOI

10.1002/hbm.22843

Peer reviewed

◆ Human Brain Mapping 36:3227–3245 (2015) ◆

Reproducibility of Brain-Cognition Relationships Using Three Cortical Surface-Based Protocols: An Exhaustive Analysis Based on Cortical Thickness

Kenia Martínez,[1,2]* Sarah K. Madsen,[3] Anand A. Joshi,[4] Shantanu H. Joshi,[5] Francisco J. Román,[1] Julio Villalon-Reina,[4] Miguel Burgaleta,[6] Sherif Karama,[7] Joost Janssen,[2,8,9] Eugenio Marinetto,[2,10] Manuel Desco,[8,10,11] Paul M. Thompson,[4] and Roberto Colom[1]

[1]Departamento de Psicología Biológica y de la Salud, Facultad De Psicología, Universidad Autónoma De Madrid, Spain
[2]Departamento de Psiquiatría del Niño y del Adolescente, Instituto De Investigación Sanitaria Hospital Gregorio Marañón, Madrid, Spain
[3]USC Mark and Mary Stevens Neuroimaging and Informatics Institute, Imaging Genetics Center, University of Southern California, Los Angeles, California
[4]Biomedical Imaging Group, University of Southern California, Los Angeles, California
[5]Department of Neurology, Ahmanson Lovelace Brain Mapping Center, University of California Los Angeles, California
[6]Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain
[7]Montreal Neurological Institute (MNI), Montreal, Canada
[8]Ciber del área de Salud Mental (CIBERSAM), Madrid, Spain
[9]Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands
[10]Departamento De Bioingeniería E Ingeniería Aeroespacial, Universidad Carlos III De Madrid, Madrid, Spain
[11]Unidad De Medicina Y Cirugía Experimental, Instituto De Investigación Sanitaria Hospital Gregorio Marañón, Madrid, Spain

Abstract: People differ in their cognitive functioning. This variability has been exhaustively examined at the behavioral, neural and genetic level to uncover the mechanisms by which some individuals are more cognitively efficient than others. Studies investigating the neural underpinnings of interindividual differences in cognition aim to establish a reliable nexus between functional/structural properties of a given brain network and higher order cognitive performance. However, these studies have produced inconsistent results, which might be partly attributed to methodological variations. In the cur-

Additional Supporting Information may be found in the online version of this article.

Authors declare there is no conflict of interest.

Contract grant sponsors: Ministerio de Ciencia e Innovación, Spain; Contract grant number: SEJ2066-07890, PSI2010-20364, BES-2011-043527, AP2006-00934; Contract grant sponsor: Ministerio de Educación, Spain; Contract grant number: AP2008-00433; Contract grant sponsor: Alianza 4 Universidades Program, Spain; Contract grant number: A4U-4-2011

*Correspondence to: Kenia Martínez; Universidad Autónoma de Madrid, 28049, Madrid, Spain. E-mail: kenia.martinez.r@gmail.com

Received for publication 21 July 2014; Revised 20 April 2015; Accepted 4 May 2015.

DOI: 10.1002/hbm.22843

Published online 28 May 2015 in Wiley Online Library (wileyonlinelibrary.com).

rent study, 82 healthy young participants underwent MRI scanning and completed a comprehensive cognitive battery including measurements of fluid, crystallized, and spatial intelligence, along with working memory capacity/executive updating, controlled attention, and processing speed. The cognitive scores were obtained by confirmatory factor analyses. T_1-weighted images were processed using three different surface-based morphometry (SBM) pipelines, varying in their degree of user intervention, for obtaining measures of cortical thickness (CT) across the brain surface. Distribution and variability of CT and CT-cognition relationships were systematically compared across pipelines and between two cognitively/demographically matched samples to overcome potential sources of variability affecting the reproducibility of findings. We demonstrated that estimation of CT was not consistent across methods. In addition, among SBM methods, there was considerable variation in the spatial pattern of CT-cognition relationships. Finally, within each SBM method, results did not replicate in matched subsamples. *Hum Brain Mapp 36:3227–3245, 2015.*

Key words: surface-based methods; cortical thickness; higher order cognition

INTRODUCTION

Individual differences exist in proficiency for reasoning, problem solving, and learning from formal and informal experience. These are crucial facets of human intelligence (Deary, 2012, Nisbett et al., 2012). Neuroimaging techniques are useful for analyzing the link between brain features and high-order cognitive performance. In 2007, the parieto-frontal integration theory (P-FIT, Jung and Haier) was published. The scientific community embraced this model, because it sought to harmonize the available findings from different neuroimaging approaches at that time. Since then, new data have often been interpreted in the context of the P-FIT model, which has helped to organize the field.

Whereas the P-FIT stressed commonalities among studies, Colom (2007) noted the great variability among the evidence summarized by Jung and Haier (2007). A very small number of discrete brain areas converged in only about half of the published studies using the same neuroimaging strategy. In structural studies assessing gray matter (GM) properties, only Brodmann areas (BAs) 39–40 and 10 reached 50% of convergence across studies. Subsequent studies suggested that results are roughly consistent with the P-FIT model. Indeed, if we overlap all reported regions from previous studies almost the entire brain would be relevant for supporting intelligent behavior. Perhaps, due in part to the extent of methodological variation in the literature, even the most consistently identified brain regions show low levels of convergence.

Several variables could explain these disparate findings across studies. Potential sources of variability are: (a) variations in the way intelligence is defined (e.g., IQ vs. *g* factor, and specific domains) and measured; (b) variations in the methods for processing MR images (for a review see Colom and Thompson, 2011; Colom et al., 2010aa), (c) variations in the quantified brain feature (structure or function), tissue (e.g., white matter [WM], gray matter [GM]) or property (e.g., volume, thickness, folding pattern, responsiveness to stimulation, WM microstructure, connectivity); and (d) variations in the tested samples (e.g., sex, age, size, lesion vs. healthy subjects or even intellectual performance).

How Intelligence in Defined and Measured

Brain imaging studies of human intelligence and related cognitive factors have used disparate measures, some of which have not taken advantage of important theoretical distinctions and empirical facts derived from the psychometric approach.

Generally speaking, human intelligence can be defined as a higher-order mental ability. Nevertheless, the intelligence construct comprises a broad set of lower-level cognitive abilities and skills, meaning that interindividual variation in performance may be due to specific combinations of lower level mental processes or cognitive tasks. Psychometric models of intelligence assume that variations in cognitive performance across different situations can be summarized by a number of basic cognitive dimensions. To characterize these dimensions, scores are obtained from diverse measures, tapping several content domains (e.g., abstract, verbal, numerical, and spatial) and processing requirements, and then analyzed through factor analysis (Abad et al., 2011). The obtained latent factors are considered as common traits underlying performance in apparently disparate tests.

The cumulative empirical evidence derived from this framework supports the view that intelligence has a hierarchical structure, with more general dimensions comprised of several cognitive abilities, which are in turn defined by specific skills (Hunt, 2011). Several models describe this structure, such as, the CHC taxonomy (McGrew, 2009) or the VPR (Verbal, Perceptual, and Image Rotation) model (Johnson and Bouchard, 2005). Overall, these models converge onto the general factor of intelligence (*g*) in addition to several cognitive abilities and specific skills. Fluid–abstract, crystallized-verbal, and spatial abilities are among the most frequently considered factors of intelligence (Carroll, 1993, 2003; Hunt, 2011).

Recently, Román et al. (2014) showed that there is a substantial variability in the GM correlates along the intelligence hierarchy, stressing the importance of properly measuring the diverse ability domains. In contrast, however, most studies rely on standardized IQ indices or single measures. The very small number of studies considering psychometric *g* and specific ability domains has used one or two tests for defining latent factors (e.g., Cole et al., 2012; Haier et al., 2009; Karama et al., 2011; Langer et al., 2012). But, as pointed out by Haier et al. (2009) "without a standard test battery selected for a known and theoretically meaningful psychometric structure, comparisons of the neuro-correlates of intelligence tests are bound to be inconsistent and difficult to interpret."

As such, the same score in a given intelligence dimension might be the result of different cognitive profiles where the specific skills involved are contributing to the observed performance in a different extend. Consequentially, the involvement of different brain regions is expected to be slightly different, depending on which lower level mental processes are contributing more to the general dimension score (for instance, "Spatial Intelligence" or "Crystallized Intelligence").Conversely, we can approach the problem from the perspective that complex traits, such as general intelligence, might be decomposed in simpler components or basic cognitive processes presumably relying on the structural and functional properties of the brain. Each of these components may be differentially recruited by different tasks, and some components may be critical for explaining variability in psychometric intelligence. Systematic comparisons of brain imaging studies reveal that cognitive functions, such as attention, working memory capacity, and processing speed, involve overlapping brain regions (Cabeza and Nyberg 2000; Colom et al., 2013; Naghavi and Nyberg 2005). Some discrete brain areas underlying these processes are also common to intelligence (Colom et al., 2007; Jonides et al., 2008). The overlap in brain regions can be interpreted as explaining, at least in part, the relationship between intelligence and cognition (e.g., working memory) at the behavioral level.

Variations in the Neuroimaging Methods and the Brain Property Studied

There is another important potential source of variability that may affect reproducibility of findings: the neuroimaging processing approach used to obtain the brain properties of interest. In the intelligence field, two main approaches have been applied to study variations in macroscopic cortical anatomy using high-resolution T_1-weighted data: Voxel-based morphometry (VBM) and Surface-based morphometry (SBM). An advantage of VBM is that it requires minimal manual intervention and can be completed relatively quickly by following the well-documented, publicly available protocols. When considering the inconsistency of results for the neural correlates of cognitive measures, the answer may lie in the details of the various processing techniques. In the case of VBM, the complexity and intersubject variability across cortical anatomy can be problematic for most standard linear and nonlinear volumetric registration algorithms used by VBM pipelines (Frost and Goebel, 2012). Not accounting for this intersubject macro-anatomical variability may weaken statistical power on group statistics, because nonidentical cortical regions are compared across subjects. This may worsen the replicability of findings in independent but comparable samples. To alleviate this loss of power due to data macro-anatomical misregistration across subjects, surface-based approaches create geometrical models of the cortex using parametric surfaces and build deformation maps on the geometric models explicitly associating corresponding cortical regions across subjects (Thompson et al., 2004). Furthermore, SBM allows us to compute several GM tissue features at the local level. These features include surface complexity, GM thickness, surface area, volume, or density. Several SBM approaches exist and each protocol differs in algorithms, parameters, and required user-intervention.

Finally, another source of variation is the type of GM property studied. Most studies have focused on the relationship between cognition and voxel-based volumetric measures using VBM algorithms. Other GM characteristics, such as cortical thickness (CT) or surface area have also been related to cognition. These different GM properties tap different cytoarchitectural aspects, which may have different underlying genetic etiology and cellular mechanisms (Panizzon et al., 2009; Winkler et al., 2010). For these reasons, findings based on disparate measures might not be directly comparable.

The current research sought to address potential sources of between-study variability by (1) using three well-known SBM protocols for estimating CT; (2) a psychometric approach for defining several psychological constructs; and (3) selecting two cognitively matched but independent samples of participants. The surface-based protocols varied in their degree of human intervention and in processing features. The cognitive dimensions of interest were: fluid intelligence, crystallized intelligence, spatial intelligence, working memory capacity, executive updating, attention, and processing speed. The samples were matched for sex, age, and cognitive performance. We assessed (1) the consistency of the three SBM methods in the estimation of thickness throughout the cortical surface; and (2) the reproducibility of the brain correlates for the measured psychological factors using the three SBM methods and the two cognitively matched samples of participants. This analytic strategy was aimed to provide tentative solutions regarding the main topic addressed here: To what extent brain-cognition relationships are robust or replicable across methods and samples?

MATERIALS AND METHODS

Participants

Four hundred and five undergraduate students completed a set of intelligence tests and cognitive tasks.

Afterward, 120 participants (60 males and 60 females), representative of the full range of test scores, were invited for MRI scanning. They completed a comprehensive questionnaire including medical, neurological, psychiatric illness, and substance abuse or conditions as exclusion criteria for MRI scanning. Written informed consent following the Helsinki guidelines was obtained from all participants. One hundred and four individuals were included in the MRI study (59 females and 45 males, mean age = 19.9, SD = 1.6, age range = 18–27; 93.3% right-handed). They received a payment of €20 for their participation.

Twenty-two participants were excluded from the SBM analyses because they failed to pass quality control in one or more of the protocols used. Therefore, the final sample for this study was comprised of 82 subjects (48 females and 34 males) with a mean age of 19.9 (SD = 1.5; minimum = 18; maximum = 27). Six were left-handed.

Image Acquisition

All images were acquired on a General Electric Signa 3T magnetic resonance (MR) scanner, using a whole-body radiofrequency coil for signal excitation, and a quadrature 8-channel coil for reception. Three-dimensional (3D) T1-weighted anatomical brain MRI scans were acquired with a spoiled gradient echo (SPGR) sequence with the following parameters: TR (repetition time) = 6.8 ms, TE (echo time) = 3.1 ms, Preparation Time = 750 ms; flip angle = 12°; 1 mm slice thickness, 288 × 288 acquisition matrix (0.8 × 0.8 × 1 mm voxel size), 512 × 512 display matrix (0.47 × 0.47 × 1 mm voxel size), 240 mm field of view and 196 images (slices) in acquisition.

Psychological Measures

Twenty-one cognitive tests and tasks were administered to measure three core intelligence domains (fluid intelligence, crystallized intelligence, spatial intelligence) and four relevant cognitive processes at different levels of complexity (from more to less complex: working memory capacity, executive updating, controlled attention, and processing speed). Following the guidelines proposed by Haier et al. (2009), all psychological constructs were estimated by three different measures that varied in processing requirements and contents.

Abstract–fluid intelligence (*Gf*) refers to reasoning and novel problem-solving ability, assessing the level of complexity that subjects can handle in situations where prior knowledge is not relevant. *Gf* was measured with the advanced progressive matrices test (Raven et al., 2004), the abstract reasoning subtests from the differential aptitude test (DAT-AR) battery (Bennett et al., 1990), and the inductive reasoning subtests from the primary mental abilities (PMA-R) battery (Thurstone, 1938).

Verbal-crystallized intelligence (*Gc*) relies on the ability to cope with academic skills and knowledge, such as reading or arithmetic. *Gc* was measured by DAT-VR (verbal reasoning), DAT-NR (numerical reasoning), and PMA-V (vocabulary).

Spatial intelligence (*Gv*) involves the construction, temporary retention, and manipulation of mental images. *Gv* was measured by DAT-SR (spatial relations), PMA-S (mental rotation), and the Rotation of Solid Figures test (Yela, 1969).

Working memory capacity (WMC) is defined as the ability to simultaneously store and process information. As such, dual memory span tasks are used to measure WMC and these were the Reading Span, Computation Span, and Dot Matrix tasks (Colom et al., 2010ab).

Updating, an executive function, is based on the online addition or subtraction of information from the working memory system. Updating was measured by the 2-Back, Keep Track, and Letter Memory tasks (Colom et al., 2008).

Controlled attention (CA) is a broad cognitive function allowing the maintenance of highly active mental representations in the presence of interference. CA was measured with verbal and numerical versions of the Flanker task, along with a spatial variant of the Simon task (Colom et al., 2010ab).

Finally, processing speed (PS) estimates the amount of information that can be processed per unit of time, often measured by reaction time, and presumably taps into the efficiency of information transfer in the brain. PS was measured with simple recognition Verbal, Numerical, and Spatial tasks (Colom et al., 2008).

The measures were administered in the same order across four sessions: intelligence tests were administered in Sessions 1 and 2, and cognitive tasks were administered in Sessions 3 and 4.

Finally, confirmatory factor analyses (CFA) were computed using AMOS 16.0.1 (Arbuckle, 2007) for testing the likelihood of the postulated measurement models: (1) three primary intelligence factors (*Gf*, *Gc*, and *Gv*) defined by their three tests, and a higher-order factor representing general intelligence (*g*); and (2) four primary correlated cognitive factors defined by their three tasks. Afterward, intelligence and cognitive latent factors were related in a structural equation modeling analysis. Maximum-Likelihood was used as method of estimation. The fit of these models was assessed by the following indices:

- Chi Square/Degrees of Freedom (CMIN/DF) ratio is considered first, as is standard, because it provides a good rule of thumb for the model fit (Jöreskog, 1993). Values of approximately 2.0 or lower show a good fit.
- Root Mean Square Error of Approximation index is sensitive to misspecification of the model. Values between 0 and 0.05 indicate a very good fit, values between 0.05 and 0.08 indicate a reasonable fit, and values greater than 0.10 indicate a poor fit (Ackerman et al., 2002; Byrne, 1998).
- Comparative fit index [CFI] (Bentler, 1990), is one of the measures least affected by sample size (Fan et al., 1999). This statistic ranges between 0.0 and 1.0 with values closer to 1.0 indicating a good fit. A cut-off

criterion of CFI $\geq$ 0.95 is recognized as indicative of a good fit (Hu and Bentler, 1999).

Second, scores for the intelligence and cognitive factors were computed to capture reliable shared variance among the specific measures of each construct. This was done using the regression imputation function of the AMOS program (Arbuckle, 2007). The latent scores obtained from the model assessing the simultaneous relationships among latent factors, as well as, the scores for the raw tests and tasks, were used in the imaging analyses.

MRI Data Processing: Surface-Based Protocols

Three different processing protocols were used to quantify CT. These protocols can be organized by the degree of user interaction and the degree to which parameters can be adjusted to optimize algorithm performance. For SBM approaches, the main difference likely lies in the registration that aligns cortical features. Landmark-based registration methods typically involve some degree of user intervention to define the sulcal and gyral landmarks, while the shape-based methods can be completely automated. The main processing steps computed by each selected protocol are described below. Note that they are organized from greater to less required user intervention.

Table I shows a comparative overview of the procedures and tools used by each protocol considered here. This table allows a rapid inspection of the methodological commonalities involved in each processing step. We choose to use the same smoothing kernel size in all pipelines. Although prior research suggests that the optimal size of smoothing filters is not the same for different morphometric measures (as it depends on the scale of the effect of interest), there is not a consensus regarding which values are most appropriate for a given application (Zhao et al., 2012). Therefore, one typical kernel size (10 mm) was used for filtering the CT estimated thought the three pipelines. To achieve this, the function "SurfStatSmooth," implemented in SurfStat, was used [http://www.math.mcgill.ca/keith/surfstat/].

Also note that CT measurements are computed using the subjects' volumetric images without rescaling. The tools and procedures used in the remaining steps are sometimes shared by two pipelines, such as MINC tools for volumetric spatial normalization and intensity normalization; algorithms included in BrainSuite software to perform brain extraction and tissue classification; and t_{link} metric to compute CT. However, pipelines fully differ in the algorithms used to reconstruct and align the cortical surfaces.

Cortical pattern matching pipeline

3D T_1-*weighted* MR images from each individual were analyzed with several manual, semiautomatic, and automatic procedures. These are detailed bellow:

1. Creation and manual editing of whole-brain and hemispheric masks for native space brain volumes. These masks were created using Brain Surface Extractor (http://brainsuite.org/processing/surfaceextraction/bse/; Shattuck et al., 2001) which is part of the BrainSuite software. Two independent expert-raters manually edited the brain masks, achieving a high inter-rater reliability (>0.95).
2. Adjustment for head position and linear transformation of brain volumes and masks into a common MNI standardized coordinate space based on the ICBM 53 dataset (Mazziotta et al., 2001) with six (no scaling) and nine parameters using MINC tools (http://www.bic.mni.mcgill.ca/software).
3. Removal of nonbrain tissue (i.e., scalp, orbits) and cerebellum, and separation of left and right hemispheres. In this step, the manually edited masks were applied to the brain volumes using MINC tools.
4. Correction for intensity nonuniformity artifacts using N3 (Sled et al., 1998).
5. Automated classification of the volumetric images in MNI space, without rescaling into GM, white matter, and cerebrospinal fluid (CSF) using partial volume classifier (PVC) (http://brainsuite.org/processing/surfaceextraction/pvc/) included in the BrainSuite software.
6. Extraction of the cortical surface based on the multiple surface deformation algorithm (MacDonald et al., 1994). This step creates deformable surface models of the cortex using as input the brain masked volumes registered to the ICBM53 template with nine parameters. The resulting cortical surfaces are represented as a high-resolution mesh of 131,072 triangulated elements spanning 65,536 surface points.
7. Manual tracing of 32 major sulcal and gyral landmarks on the lateral and medial surfaces of each hemisphere using an interactive software (MNI Display; MINC tools) and a standardized sulcal labeling protocol [http://resource.loni.usc.edu/resources/downloads/research-protocols/sulcal-anatomy/].
   Two trained experts independently traced each of the 14 sulci on the lateral brain surface [Sylvian fissure; central, precentral, postcentral, superior temporal sulcus (STS) main body; STS ascending branch; inferior temporal; superior frontal; inferior frontal; intraparietal; transverse occipital; olfactory; occipitotemporal; and collateral sulci] in each hemisphere on the surface rendering of each subject's brain. An additional set of 11 sulci was outlined on each medial surface (callosal sulcus, inferior callosal outline, superior rostral sulcus, inferior rostral sulcus, paracentral sulcus, anterior and posterior segments of the cingulate sulcus, parieto-occipital sulcus, anterior and posterior segments of the calcarine sulcus, and the subparietal sulcus). In addition to

TABLE I. Main surface-based processing steps and tools used in each protocol considered

	CIVET	BrainSuite	CPM
Processing steps	Procedure/Software	Procedure/Software	Procedure/Software
Volumetric spatial normalization	Linear registration to ICBM 152 with nine parameters using MINC tools	Linear registration to Colin27 with six parameters using FSL tools	Linear registration to ICBM 53 with nine parameters using MINC tools
Intensity normalization	Nonparametric Nonuniform Intensity Normalization (N3; Grayscale level-based methods) using MINC tools	Bias Field Corrector (BFC; Grayscale level-based methods) implemented in BrainSuite software	Nonparametric Nonuniform Intensity Normalization (N3; Grayscale level-based methods) using MINC tools
Brain extraction	Brain Extraction Tool (BET) (edge-based method) using FSL tools	Brain Surface Extractor (BSE; Region-based method) implemented in BrainSuite software. Manual editing of the brain mask to improve the brain extraction	Brain Surface Extractor (BSE; Region-based method) implemented in BrainSuite software. Manual editing of the brain mask to improve the brain extraction
Tissue classification	Intensity Normalized Stereotaxic Environment for Classification of Tissues (INSECT; intensity-based method) implemented in CIVET pipeline	Partial Volume Classifier (PVC; intensity-based method) implemented in BrainSuite software	Partial Volume Classifier (PVC; intensity-based method) implemented in BrainSuite software
Cortical surface reconstruction	Constrained Laplacian-based ASP (CLASP) implemented in CIVET pipeline. The resulting hemispheric mesh has 81924 triangles and 40962 vertices for all subjects	Cortical Surface Extraction algorithm implemented in BrainSuite software. The resulting whole brain mesh has approximately 500,000 vertices depending on the subject's brain	Multiple Surface Deformation (MSD). The resulting hemispheric mesh has 131072 triangles and 65536 vertices for all subjects
Cortical surface registration	Nonlinear hierarchical deformable registration to a high-resolution average template based on depth potential function (automatic shape-based method) implemented in CIVET pipeline.	Nonlinear registration to a high-resolution template based on curvature (automatic shape-based method) implemented as SVREG toolbox in BrainSuite software. Includes an iterative refinement step to improve feature correspondence based on landmarks (ROIs and sulci) transferred from the template to the subjects' surfaces during registration.	Diffeomorphic sulcal shape matching across subjects (landmark-based method based in manually delineated major cortical sulci)
Cortical thickness measurement	t_{link} metric	t_{link} metric	3D Eikonal equation
Smoothing kernel size	FWHM = 10 mm (for all morphometric measures)	FWHM = 10 mm (for all morphometric measures)	FWHM = 10 mm (for cortical thickness)

contouring the major sulci, a set of seven midline landmark curves bordering the longitudinal fissure was outlined in each hemisphere to establish the hemispheric gyral limits. The spatially registered gray-scale image volumes in coronal, axial, and sagittal planes were available simultaneously to help disambiguate brain anatomy. The detailed criteria for delineating the cortical lines and for starting and stopping points for each sulcus are provided in the protocol. The tracer reliability was measured using the 3D root mean square difference (in millimeters) between sulci in a set of six test brains and those of a gold standard set. Disparities between the test and gold standard brains were computed to be <2 mm for all landmarks (Sowell et al., 2002).

8. Nonlinear elastic warping between meshes using a diffeomorphic sulcal shape atlas computed on the tested sample (Joshi et al., 2012c). This step is run separately for left and right sulcal landmarks, and distorts the anatomy of one subject into another, matching sulcal

features. The final outputs are resampled meshes with vertices in anatomical homology.

9. Gray matter thickness was calculated using the Eikonal Fire equation (Sapiro, 2001). Although the brain image volumes acquired for this study had voxel dimensions of 0.47 × 0.47 × 1 mm, the image data from step 5 (i.e. the segmented volume image in MNI coordinates without rescaling) were supersampled to create voxel dimensions of 0.33 mm cubed; thus the 3D Eikonal equation was applied only to subsampled voxels segmented as GM (Sowell et al., 2004). Next, to map GM thickness onto the surface rendering of each subject, the coordinate for each brain surface point (anatomically matched across individuals) was mapped to the same anatomical location in their thickness volume and a smoothing kernel was used to average GM thickness within a 10 mm sphere at each cortical surface point.
10. Blurring of each subject's CT map using a 10-mm full width at half maximum (FWHM) surface-based diffusion smoothing kernel.

BrainSuite pipeline

3D T_1-*weighted* MR images were mainly processed by BrainSuite13's surface extraction and registration tools [http://neuroimage.usc.edu/neuro/BrainSuite]. Some independent preprocessing tools were also used, as detailed below:

1. Whole-brain masks were created and manual edited for native space brain volumes using the procedures and tools described for CPM pipeline (1).
2. Removal of nonbrain tissue (i.e., scalp, orbits) and cerebellum. In this step, the manually edited masks were applied to brain volumes in native space using FSL tools [http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Fslutils#Tools/].
3. Adjustment for head position and linear transformation of brain volumes and masks into a common MNI standardized coordinate space based on the Colin 27 template (Holmes et al., 1998) with six parameters (no scaling) using FSL tools.
4. Correction for intensity nonuniformity artifacts using the Bias Field Corrector from BrainSuite.
5. Automated classification of volumetric images in MNI space without rescaling into GM, WM, and CSF using the PVC included in the BrainSuite software.
6. Extraction of the cortical surfaces. To achieve this, classified images containing only the cerebrum were submitted to the surface extraction steps included in BrainSuite. Before the inner surface generation, a cleaned and corrected mask of the inner cortical boundary must be obtained by running four tools: inner cortex mask selection and scrubbing, topological correction, and wisp removal. Then, a mesh model representing the boundary between WM and cortical GM is generated. Next, the pial surface is generated by an iterative process, which dilates the inner surface until the GM-CSF boundary is reached. The result is a one-to-one map between the points on the inner cortical surface model and the pial surface model, both containing approximately 500,000 vertices depending on the subject. Finally, the surfaces are split into left and right hemispheres.
7. Nonlinear registration of the cortical surfaces for each hemisphere to a high-resolution labeled surface template using the SVREG tool (Joshi et al., 2012b) from the BrainSuite package. This spatial alignment step is performed on a medial cortical surface mesh, which is computed based on the inner and outer boundaries for the atlas and the subjects. In the atlas used as target, 35 cortical regions of interest and 26 sulci are delineated and transferred to each subject's surface during the registration procedure. First, the subjects' surfaces are aligned to the atlas by matching a "flattened" version of mean curvature maps computed for the atlas and the subjects, where the sulci fundi are represented with negative values and gyral crowns with positive values. After performing the atlas-to-subject registration, the sulcal curves and cortical labels from the atlas are applied to the subjects' cortical surfaces. Next, a refinement of the labels and sulcal curves is performed to improve the spatial accuracy of imposed landmarks from atlas to subjects' surfaces. The meshes resulting from these transformations are expected to ensure the one-to-one correspondence of anatomical features across subjects at each cortical point, having each surface model 312,132 triangulated elements (156,188 vertices) for the left hemisphere and 314,276 triangles (157,260 cortical points) for the right hemisphere.
8. Computation of CT at each cortical point using the t_{link} metric (Lerch and Evans, 2005).
9. Blurring of each subject's CT map using a 10-mm FWHM surface-based diffusion smoothing kernel.

CIVET pipeline

3D T_1-*weighted* MR images were submitted to the CIVET image-processing environment (version 1.1.9) developed at the MNI, a fully automated pipeline to extract and coregister the cortical surfaces for each subject (Ad-Dab'bagh et al., 2006). The main pipeline processing steps include:

1. Correction for intensity nonuniformity artifacts using N3 (Sled et al., 1998).
2. Linearly registration of native (i.e., original) MR images to standardized MNI-Talairach space, based on the ICBM152 data set (Collins et al., 1994; Mazziotta et al., 1995; Talairach and Tournoux, 1980)

using nine parameters (three rotations, three translations, three scales) and using MINC tools.
3. Brain masking using an improved version of BET from FSL (Smith, 2002).
4. Tissue classification into GM, WM, CSF, and background using a neural net classifier [INSECT] (Zijdenbos et al., 2002). Then, partial volume fractions of these tissue types were computed for each brain voxel (Kim et al., 2005; Tohka et al., 2004).
5. Image fitting with a deformable mesh model to extract inner (WM/GM interface) and outer (pial) cortical surfaces for each hemisphere with the third edition of CLASP. This produces high-resolution hemispheric inner and outer surfaces (having one-to-one correspondence) with 81,924 polygons each (40,962 vertices or cortical points per hemisphere) (Kim et al., 2005; MacDonald, 1998; MacDonald et al., 1994, 2000).
6. Nonlinear registration of left and right hemisphere cortical surfaces to a high-resolution average surface template iteratively generated from the ICBM152 data set using a depth-potential function (Boucher et al., 2009; Lyttelton et al., 2007) to establish intersubject correspondence of the cortical features.
7. Rescaling of the aligned cortical surfaces back to native space dimension using the inverse of the scaling parameters of the corresponding linear volumetric transformation matrix (obtained in step 2). Thus, CT measurement was made in native space. This avoids having GM morphometric measurement biased by the scaling factor introduced by the linear transformations applied to each subject's brain in step 2.
8. Computation of CT at each cortical point using the t_{link} metric (Lerch and Evans, 2005).
9. Blurring of each subject's CT map using a 10-mm FWHM surface-based diffusion smoothing kernel.

Statistical Analysis

Statistical analyses were implemented using SurfStat, a statistical toolbox (Worsley et al., 2004) created for MATLAB 7 (The MathWorks) [http://www.math.mcgill.ca/keith/surfstat/] at the MNI.

First, the distribution and variability of CT across the cortex was analyzed by computing the mean and standard deviation at each cortical point (vertex) in the complete sample ($n = 82$). The analysis was repeated using two subsamples ($n = 41$) matched for sex, age and all the cognitive factors considered. Also, we generated 100 additional pairs of matched subsamples to assess whether the observed patterns in the two original matched subsamples could be replicated. The process for generating these subsamples was based on computing the Euclidean distance among subjects' cognitive and age scores while ensuring the same size and similar characteristics as the original matched subsamples. The imposed characteristics were: (1) equal proportion of males and females; (2) nonsignificant differences between subsamples in age, and in the general estimates of cognition (represented by the latent factors) at $\alpha = 0.05$. Supporting Information Figure S1 shows the distribution of each cognitive factor in the 100 pairs of matched subsamples.

Next, *Student*'s *t*, *Pearson*'s *r* and *P*-values maps were obtained and visualized in MATLAB for the main effect of each first order latent factor/psychological measure over CT after controlling for sex, age, and handedness. Pearson's *r*-values were obtained by transforming the *Student*'s *t* values according to the following formula: $r = t/\text{sqrt}\ (t^2 + \text{df})$; where sqrt = square root and df = degree of freedom.

The uncorrected statistical maps allowed us to visually inspect the spatial patterns of brain-behavior relationship across morphometry protocols, cognitive constructs, and samples. Nevertheless, the resulting *t*-maps were corrected for multiple comparisons (α =0.05) via the false discovery rate (FDR) method (Benjamini and Hochberg, 1995; Genovese et al., 2002).

We followed Cohen (1988) for interpreting the magnitude of the observed effect sizes (Pearson's *r*-values): correlations greater than 0.5 were considered large, 0.5–0.3 moderate, and 0.3–0.1 small. We used G*Power software to perform an a priori and post hoc power analysis. This helped us to interpret the relationship between sample size, effect size, and threshold for significance. Although there are no formal standards for power, 0.80 was used as cutoff for adequacy.

RESULTS

Psychological Measures and Constructs of Interest

Raw data derived from the intelligence tests and cognitive tasks

The raw correlations and descriptive statistics (Mean and Standard Deviation [SD]) for the intelligence tests and cognitive tasks computed for the complete ($n = 104$) and final ($n = 82$) samples are reported in Table SI (Supporting Information). This correlation matrix was submitted to a CFA to test the postulated measurement models.

Confirmatory Factor Analysis

Intelligence tests. First, the following confirmatory model was tested: (1) fluid-abstract intelligence (*Gf*) was defined by the Raven advanced progressive matrices test, the inductive reasoning subtest from the PMA (PMA-R), and the abstract reasoning subtest from the DAT (DAT-AR). (2) Crystallized-verbal intelligence (*Gc*) was defined by the vocabulary subtests from the PMA (PMA-V), the verbal reasoning subtest from the DAT (DAT-VR), and the numerical reasoning subtest from the DAT (DAT-NR). (3)

Spatial intelligence (*Gv*) was defined by the rotation of solid figures test, the mental rotation subtest from the PMA (PMA-S), and the spatial relations subtest from the DAT (DAT-SR). Further, a higher-order factor representing general intelligence (*g*) was also defined. The fit for this model and the structural weights are depicted in Figure S2 (Supporting Information). Note that fluid intelligence (*Gf*) was the primary factor best predicted by the higher-order factor (*g*). Indeed, the measurement model shows that *Gf* is perfectly predicted by *g* (factor loading = 0.99). Then, a model where three intercorrelated latent factors (*Gf*, *Gc*, and *Gv*) were defined by their respective measures (Supporting Information Fig. S3) was tested. It must be noted that both models were equivalent in terms of model fit, but the model without the *g*-factor is more parsimonious because it includes a lower number of latent factors.

Cognitive tasks. Next, the measurement model for the cognitive constructs was tested. WMC was defined by the Reading span, Computation span, and Dot Matrix tasks; executive updating (UPD) was defined by the 2 Back, Letter Memory, and Keep Track tasks; PS was defined by verbal, numerical, and spatial short-term recognition speed tasks; and controlled attention was defined by the verbal and numerical flanker tasks, along with the Simon task (See Supporting Information Fig. S4). Results showed that the latent factors for WMC and UPD were perfectly correlated, so they were collapsed into the same factor, namely WMC. Indices in Fig. S5 (Supporting Information) show that this last model has excellent fit.

Simultaneous relationship among the assessed psychological constructs. Figure 1 depicts the final model showing correlations among the resulting six latent factors—fluid intelligence (*Gf*), crystallized intelligence (*Gc*), spatial intelligence (*Gv*), WMC, PS, and controlled attention (CA)—along with the regression weights for the specific measures attached to each factor. Note that this model has appropriate and very similar fit indices, when computed using both the original sample ($n = 104$) and the final sample ($n = 82$, blue numbers in Fig. 1). Results suggest that, considering the simultaneous relationships among all intelligence and cognitive factors, at $\alpha = 0.01$, CA and PS were unrelated to the three intelligence factors, whereas WMC was related to all of them in both samples. However, at $\alpha = 0.05$, *Gf*, *Gc*, and *Gv* were correlated with PS (orange lines) and *Gc* was correlated with CA in the larger sample only. In the final sample ($n = 82$), only the relationship between *Gf* and PS remained significant at $\alpha = 0.05$.

Obtaining intelligence and cognitive scores. Finally, general scores for the intelligence and cognitive factors (from model in Fig. 1) were computed for capturing reliable shared variance among the specific measures of each construct, taking into account the simultaneous relationship among all the variables within the model. Correlations among the resulting imputed latent scores are also shown in Table SII (Supporting Information). As can be observed, correlations among the resulting imputed latent scores are higher, because the imputation method used to estimate latent scores takes into account the simultaneous relationships among all the variables included within the model.

Afterward, statistical analyses assessed the relationship between CT and the six latent intelligence and cognitive factor scores.

Neuro-Anatomical Networks for Intelligence and Cognition: Consistency Across Surface-Based Methods and Samples

The results are divided into two main sections. The aim of the first section is to provide tentative answers regarding the impact of neuroimaging processing protocols on the observed findings. To accomplish this, three surface-based protocols (Cortical Pattern Matching [CPM], BrainSuite and CIVET pipelines) were used to obtain CT throughout the cortex for all subjects. Thereafter, we obtained, using the outputs from these protocols, (1) the distribution and variability of CT across the cortex; and (2) the brain correlates for each psychological latent factor. Finally, the results from these analyses were compared in terms of commonalities.

The second section deals with the heterogeneity promoted by the sample tested. First, the entire sample was divided into two matched subsamples in terms of sociodemographic variables, cognitive performance, and size. Then, the same statistical analyses computed for the whole sample were performed for both subsamples using the CT measurement derived from the three surface-based protocols. Results are discussed in terms of (a) convergence between the samples in the distribution and variability of CT computed across the cortex and (b) the commonalities across matched samples in their CT-cognition correlates. Finally, these analyses are repeated in 100 additional pairs of matched subsamples for testing whether the patterns found in the two original matched subsamples could be replicated.

Pearson r-maps were computed for testing the main effects of the psychological constructs on the CT after controlling for age, sex, and handedness and were color-coded on an average surface template generated for each SBM protocol used. In the figures representing brain surfaces in this manuscript, we will use the convention that sagittal views of right hemisphere (frontal pole pointing to the readers' right) and left hemisphere (frontal pole pointing to the readers' left).

The reported results are uncorrected, as we failed to find significant brain-behavior associations after correcting for multiple comparisons via FDR. Thus, we will present significant findings at three standard uncorrected thresholds (0.005, 0.001, and 0.0001). We did not report vertex coordinates in any standard space (e.g., MNI or Talairach) for peak values, because the extracted surfaces from the three SBM protocols are different and, therefore, do not have

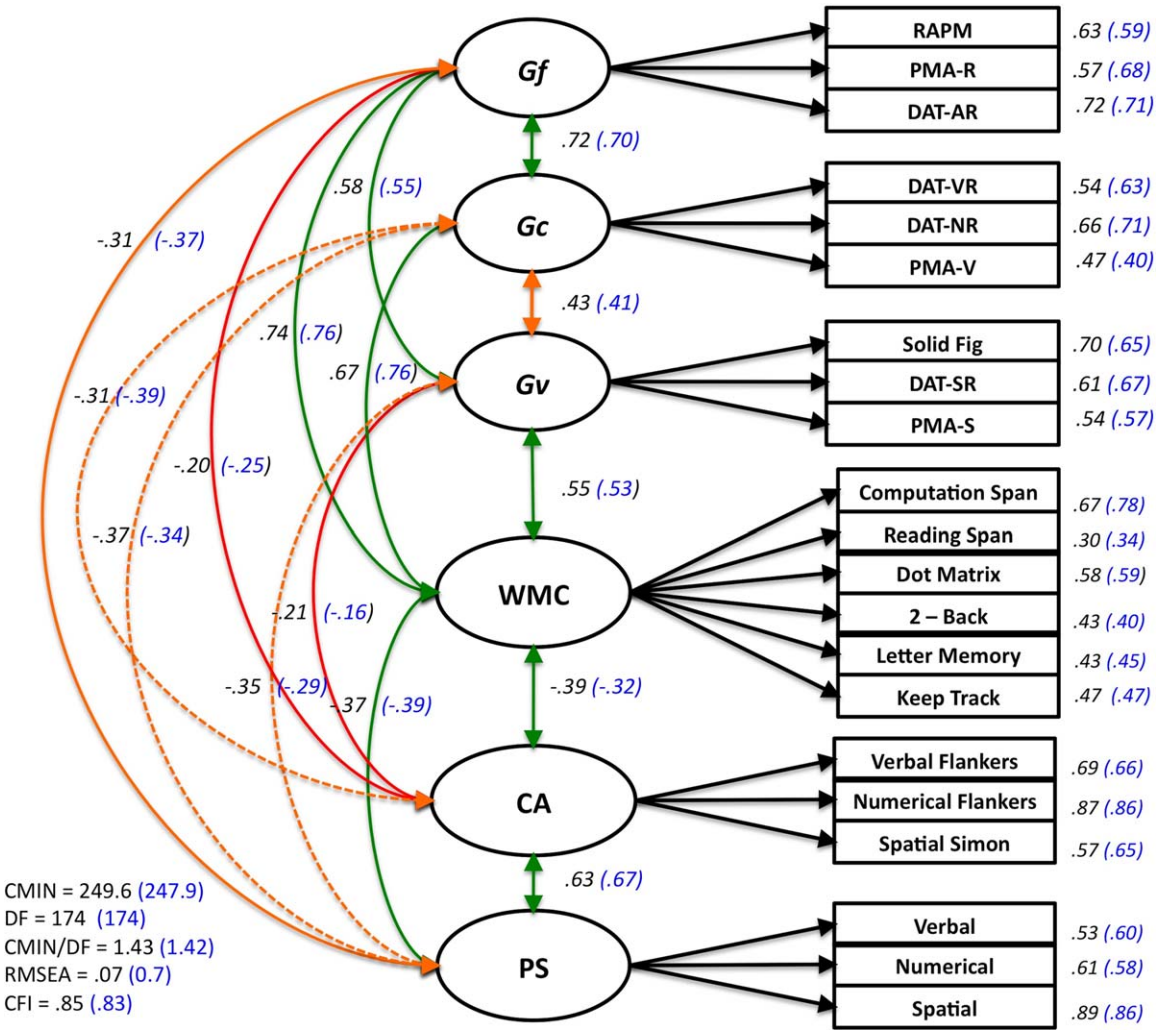

Figure 1.

Final confirmatory factor model for the intelligence and cognitive measures. Blue weights and correlations (in parenthesis) correspond to those computed for the final sample ($n = 82$). Green and orange lines represent significant correlations for both sample sizes at $\alpha = 0.01$ and $\alpha = 0.05$, respectively. Orange dash lines represents correlations that are significant at $\alpha = 0.05$ for the larger sample but not for the final one ($n = 82$). Red lines denote nonsignificant correlations. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

exact correspondence. The standard coordinate systems are based on volumetric data. Mapping vertices in mismatched meshes to voxels does not guarantee that a given vertex will fall in the same coordinate, but may be close depending on the difference between meshes' resolution. So, we decided to visually inspect the major anatomical boundaries in each mesh and report patterns of overlap.

Validation of Findings Using Different Surface-Based Protocols

Convergence in the distribution and variability of cortical thickness estimated using different protocols. As shown in Figure 2, the distribution (mean) and variability (SD) of CT throughout the cortex for the whole sample changed depending on the surface-based protocol used to compute this morphometric measure.

Regarding the distribution of CT (Fig. 2, Top), using the CIVET protocol, the highest values were found in the insular cortex, the medial temporal pole/entorhinal, and the posterior portion of the medial orbitofrontal gyrus. Using BrainSuite, the highest values were found mainly in the anterior cingulate, medial superior frontal gyrus, and anterior portion of lateral superior temporal gyrus, medial temporal pole/enthorinal/anterior parahippocampal gyrus, and insular cortex. In both methods, the thinnest areas were

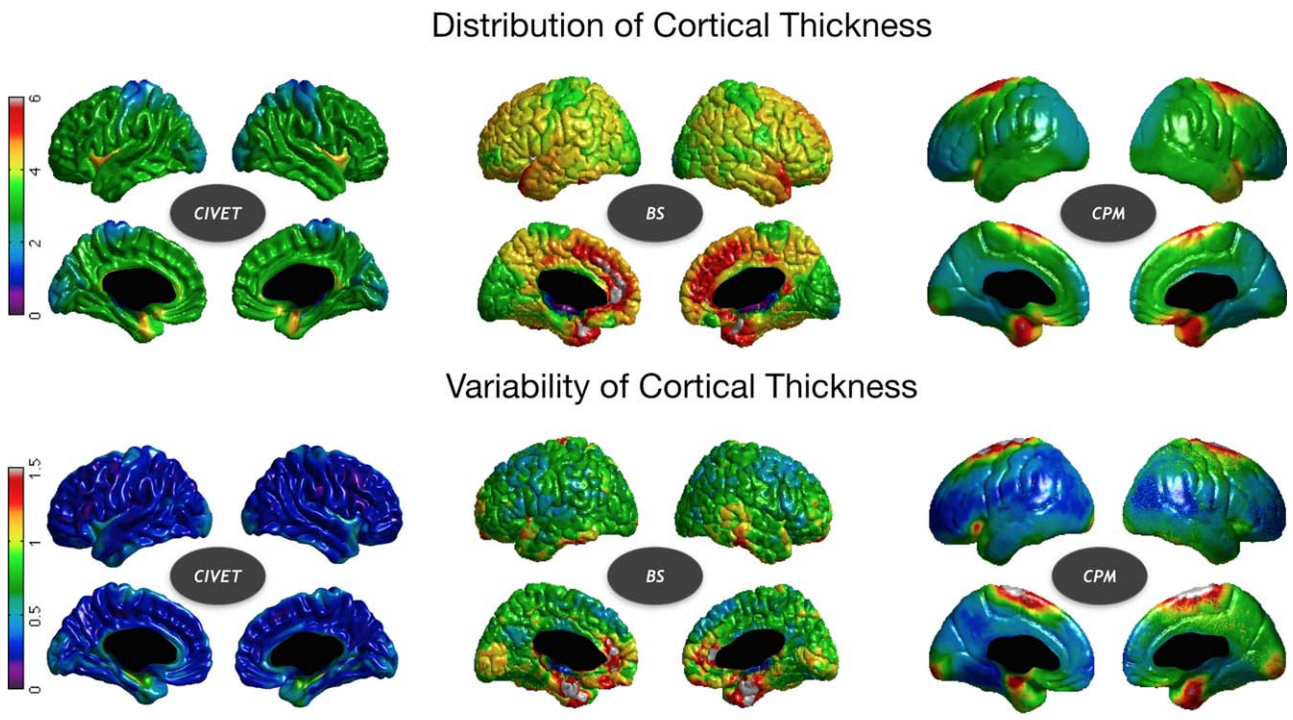

Figure 2.
Distribution and variability of CT computed through different surface-based protocols: CIVET, BrainSuite (BS) and CPM. Figure shows mean values (Top) and standard deviation values (Bottom) at each vertex. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

observed in lateral and medial occipital, and in the lateral superior and medial portions of the precentral and postcentral gyri. Finally, the distribution of CT calculated by the CPM pipeline differed from both CIVET and BrainSuite, except in some occipital regions, insular cortex, and the medial temporal pole, where all methods converge.

Standard deviation maps shown in Figure 2 (Bottom) suggested that each surface-based protocol influences the observed pattern of variability. Note that values obtained from the CIVET pipeline were more homogeneous and closer to zero as compared to the other two protocols.

Overlaps between protocols in those brain correlates computed for each psychological latent factor. As shown in Supporting Information Figures S6 and S7, there was a poor convergence among methods in the pattern of associations between morphological measures and cognitive factors across the cortex: significant results did not overlap for any latent factor.

An example of this pattern is displayed in Figure 3, where significant results at three commonly used uncorrected statistical thresholds for three psychological latent factors at different levels of cognitive complexity are noted. Higher *r*-values for the relationship between each of the represented constructs and CT were widespread in the surfaces depending on the protocol used. Furthermore, in some cases, equivalent brain regions appear to exhibit opposite patterns in their relationship with the psychological factors. For instance, vertices located in the right superior parietal cortex correlated with *Gv* using the BrainSuite and CPM protocols, but in opposite directions (see Supporting Information Fig. S5).

Validation of Findings Considering Cognitively Matched Samples

Assessment of the equivalency between samples in age, sex, and cognitive performance. Descriptive results (Mean and SD) for the psychological latent constructs and raw measures (intelligence tests and cognitive tasks) after dividing the complete sample into two matched groups of subjects are shown in the Supporting Information (Table SIII). Each of these subsamples comprises 41 individuals matched by sex (24 females and 17 males each) and age (subsample A: mean = 19.8 and SD = 1.76; subsample B: mean = 19.9 and SD = 1.31; *t*-test for equality of means = −0.43; $P = 0.67$, equality of variance assumed). *t* and *P* values for Independent Samples Tests and effect sizes (Cohen's *d*) are also reported. For all instances, the Levene's Tests allowed to assume equality of variances between subsamples ($P < 0.05$). There were not significant differences between subsamples in the general estimates of cognition (represented by the latent factors) at $\alpha = 0.05$.

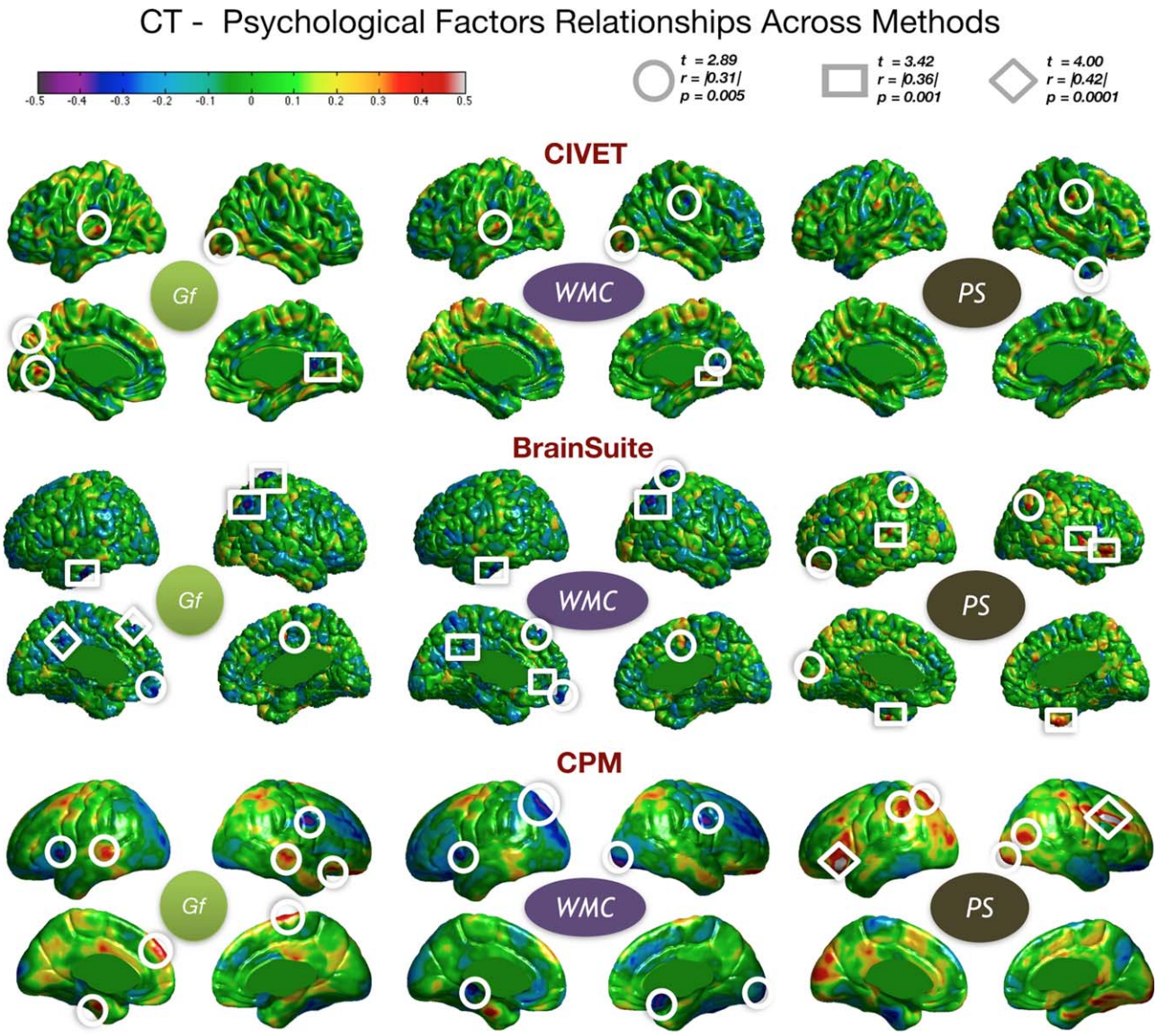

Figure 3.

r-maps for the complete group ($n = 82$) representing the magnitude of the relationship between CT obtained by three different surface-based protocols and three psychological latent factors at different levels of cognitive complexity. *r*-values for three commonly used uncorrected statistical thresholds and their corresponding *t*-values are displayed on the left. Circles, squares and rhombus indicate *r*-values significant at $p < 0.005$, $p < 0.001$, and $p < 0.0001$, respectively. *Gf*, fluid intelligence; WMC, working memory capacity; PS, processing speed. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Convergence in the distribution and variability of cortical thickness in two cognitively matched samples using different Surface-based protocols. Average CT (Supporting Information Fig. S8) did not differ between subsamples at any cortical point within the three surface-based protocols (Independent Samples *t*-test, $\alpha = 0.05$ FDR corrected), which corresponds to the same pattern and disparity between neuroimaging methods observed for the whole sample (see an example of this pattern in Fig. 4, Top). Moreover, correlations between mean values in subsample A and subsample B were very high (close to one) for the three pipelines (CIVET: $r = 0.99$; BrainSuite: $r = 0.97$; CPM: $r = 0.99$).

This trend is also applicable to standard deviation maps (Supporting Information Fig. S9), as almost for all vertices the equality of variances between subsamples is assumed (Levene's Test, $\alpha = 0.05$ FDR corrected). In a few regions, one subsample had less variability in CT than the other subsample (proportion over the total number of vertices:

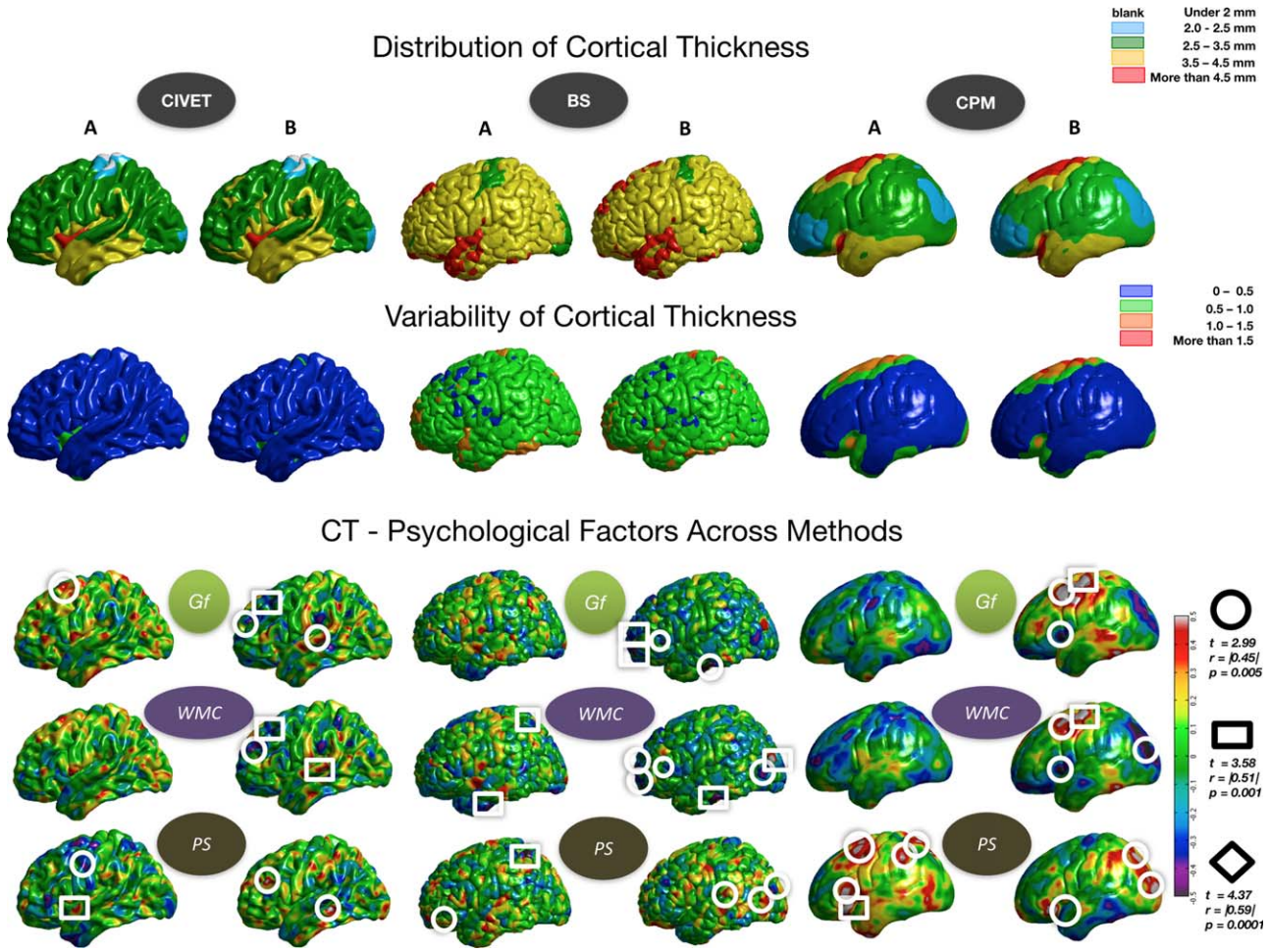

Figure 4.

Mean and standard deviation maps for CT (Top and Middle rows, respectively) computed using different surface-based protocols (CIVET, BrainSuite, and CPM) in two cognitively matched subsamples (**A** and **B**). *r*-maps representing the magnitude of the relationship between CT and three psychological latent factors at different levels of cognitive complexity are displayed for both subsamples and the pipelines used (Bottom row). Results shown are for the lateral left hemisphere. Circles, squares, and rhombus indicate *r*-values significant at $p < 0.005$, $p < 0.001$, and $p < 0.0001$, respectively. *Gf*, fluid intelligence; WMC, working memory capacity; PS, processing speed. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

CIVET = 0.005; BrainSuite = 0.014; CPM = 0). Notably, the specific spatial location of these results was highly dependent on the surface-based method used to process the 3D MR images (see an example of this pattern in Fig. 4, Middle). Finally, correlations between standard deviation values in both subsamples were slightly lower compared with those obtained for the distribution of CT, but were still high (CIVET: $r = 0.88$; BrainSuite: $r = 0.74$; CPM = 0.90).

These results were replicated when assessing the 100 pairs of matched subsamples as illustrated in Supporting Information Tables SIV and SV.

Overlap between samples in those brain correlates computed for each psychological latent factor using different surface-based protocols. As shown in Figures S10 and S11 (Supporting Information) convergence between the subsamples was poor across SMB methods. Significant associations between CT and psychological factors across the cortex did not overlap. *r*-values were close to zero for correlations between *t*-maps resulting from each subsample, protocol, and psychological measure (see Supporting Information Tables SVI and SVII). A lack of convergence was also found in the 100 pairs of matched subsamples (see Supporting Information Fig. S12 and Table SVIII). For

instance, there was a clear disparity between subsamples A and B for the brain correlates of fluid intelligence (*Gf*), WMC and PS, shown in the lateral view of the left hemisphere in Figure 4, Bottom.

The most prominent findings observed for the complete sample were a combination of those obtained for both subsamples, regardless of the method used. Figure 5 shows an example of this pattern, when spatial intelligence (*Gv*) and CT computed by the CPM processing pipeline are closely inspected. Significant findings for the whole sample were almost exclusively driven by those found for subsample B. Furthermore, when specific regions, such as those highlighted by green arrows, were related in the same direction for the subsamples, they tend to appear as relevant or became significant for the complete sample (e.g., left insular, inferior portion of right precentral gyrus).

Nevertheless, there were several cortical points where the relationship between *Gv* and CT was significant for both subsamples, but in opposite directions (yellow circles). In those instances, as highlighted by yellow arrows in the *r*-maps for the whole sample, the relationship between this psychological factor and CT became unsubstantial or null in the larger sample. These tendencies extended to other psychological factors.

A basic a priori power analysis demonstrated that with a sample size close to 140 the minimum detectable effect size at $\alpha = 0.005$ with a 0.8 1-β is 0.3 (moderate effect). To detect effect sizes of 0.5 or larger only 45 subjects are required. In our data, a post hoc power calculation revealed that the probability of detecting a significant moderate effect (0.3) in the whole sample ($n = 82$) at an alpha threshold of 0.005 was 49%, whereas this probability was reduced until 19% for the subsamples ($n = 41$). Large effects (0.5) were detectable in 98% of cases for the whole sample and 71% for the smaller sample. Thus, only large effect sizes are expected to be consistently detectable with our current sample sizes ($n = 82$ and $n = 41$). In our data, the Type II error increases considerably for moderate and small effects.

Our results showed that the peak correlation values for significant findings at a certain α level and spatial location were very different for both sample sizes. Compared with the larger sample, the smaller samples showed higher effect sizes. As shown in Figure 5, the positive findings in the left middle frontal and right superior parietal region for the whole sample ($n = 82$) are moderate (equal or less than 0.35). For subsample B ($n = 41$) these values were higher than 0.50. Post hoc power calculation directly relies on sample size, but also on the effect size and the threshold for significance. The statistical power of findings in these regions at $\alpha = 0.005$ for the smaller subsample B is close to 0.71, while when the sample is enlarged, the effect size is reduced and the statistical power decreases to approximately 0.5.

In our study, an increase in sample size while keeping the probability of Type I error constant, did not guarantee improved post hoc power, presumably due to the fact that adding cases to the sample decreased some of the large effect sizes. As mentioned previously, significant findings for the complete sample were a composite of those obtained from both subsamples. In the case of *Gv* and CT computed by the CPM pipeline, findings for the whole sample were almost exclusively driven by those discovered for subsample B, the *t*-values being close to zero in subsample A or even correlating in the opposite direction for certain cortical regions.

DISCUSSION

Here, we have systematically analyzed the structural brain correlates of a representative set of intelligence and cognitive factors (fluid intelligence, crystallized intelligence, spatial intelligence, WMC/executive updating, CA, and PS). Our goal was to overcome potential sources of variability: surface-based structural neuroimaging protocols, cortical morphological measurement, the nature of the cognitive measurements, and sample characteristics. Our main interest was focused on revealing tentative explanations about why we still do not have reliable and reproducible brain networks supporting different facets of cognition.

How Surface-Based Protocols May Influence Convergence Across Studies?

We failed to find consistency among results derived from the different surface-based imaging protocols used.

First, and especially important, the CT estimation and its variability at each cortical point were quite different across surface-based protocols.

Second, there was only a small convergence among methods in the pattern of associations observed between CT and cognitive factors across the cortex: the vertices with peak values were located in nonequivalent anatomical regions. This observation is consistent with the low convergence observed in the Jung and Haier's (2007) review, as well as, in the research reports published afterward (Colom et al. 2009). The explanation for this small convergence may have a number of different sources and we sought to identify some of them in the current study.

There were important differences in the CT maps obtained from the three SBM pipelines: neither the distribution nor intersubject variability was the same. This precludes finding substantial overlapping across pipelines in the analyzed brain-cognition relationships, as CT cannot be seen as a reliable measure in the current study. Unfortunately, there is still no available ground truth data for CT at high resolution, making it difficult to assess the biological plausibility of the outputs. We used the von Economo's postmortem histological maps as external criteria for validation. Figure 6 displays the distribution of thickness along the cortex, as well as, the variability across subjects, proposed by von Economo (1929). Visual

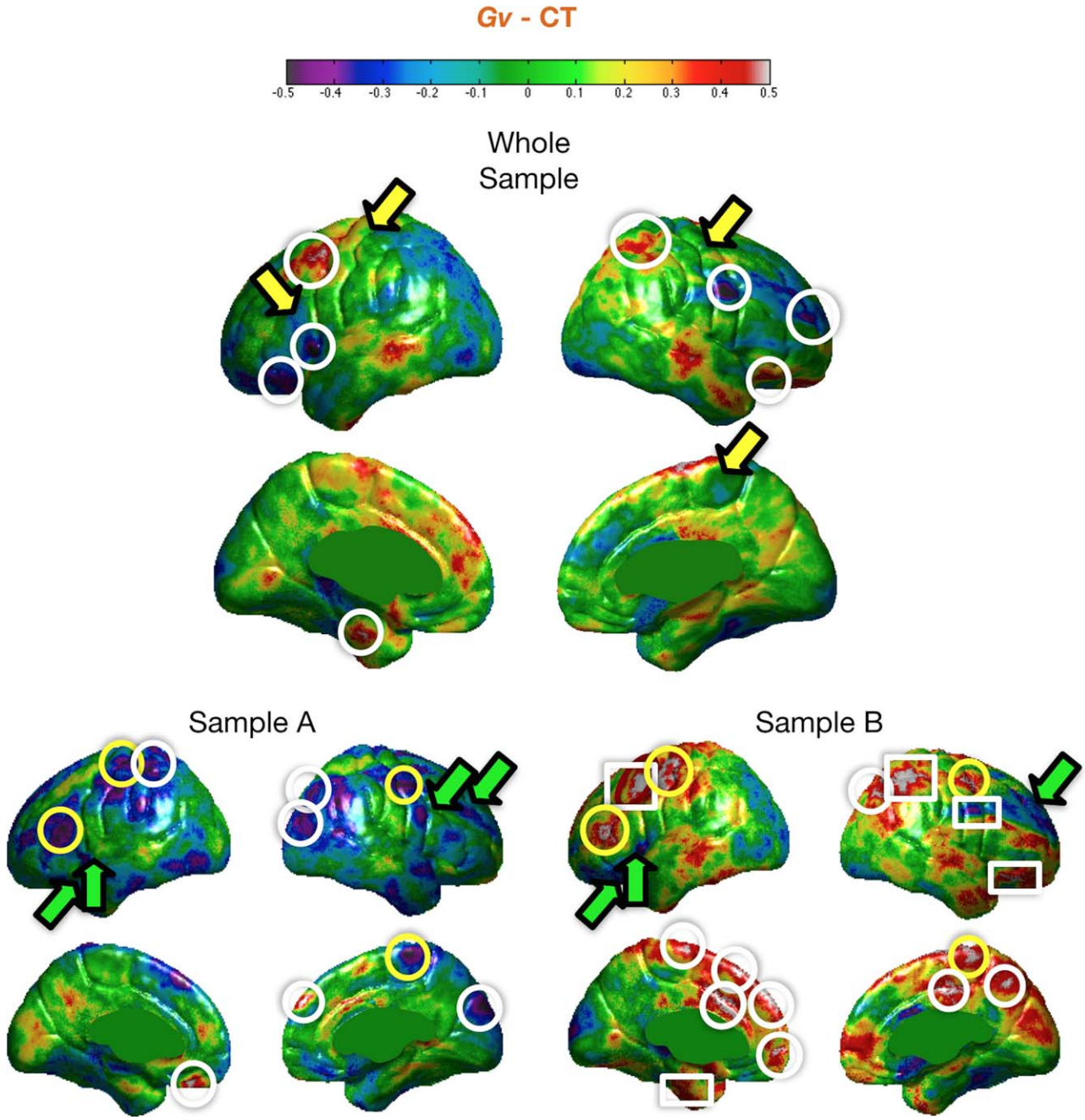

Figure 5.

r-maps for the whole sample (Top) and two cognitively matched subsamples ($n = 41$; Bottom) where is tested the main effect of spatial intelligence (*Gv*) on CT for CPM pipeline. Circles and squares highlight brain correlates significant at α = 0.005 and $\alpha = 0.001$, respectively. Yellow circles indicate brains regions where CT and *Gv* are related but in opposite directions, disappearing when the whole sample is considered (yellow arrows). Green arrows indicate broad brain regions where the relationship has the same directionality in both subsamples but without reaching significance unless the whole sample is considered. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

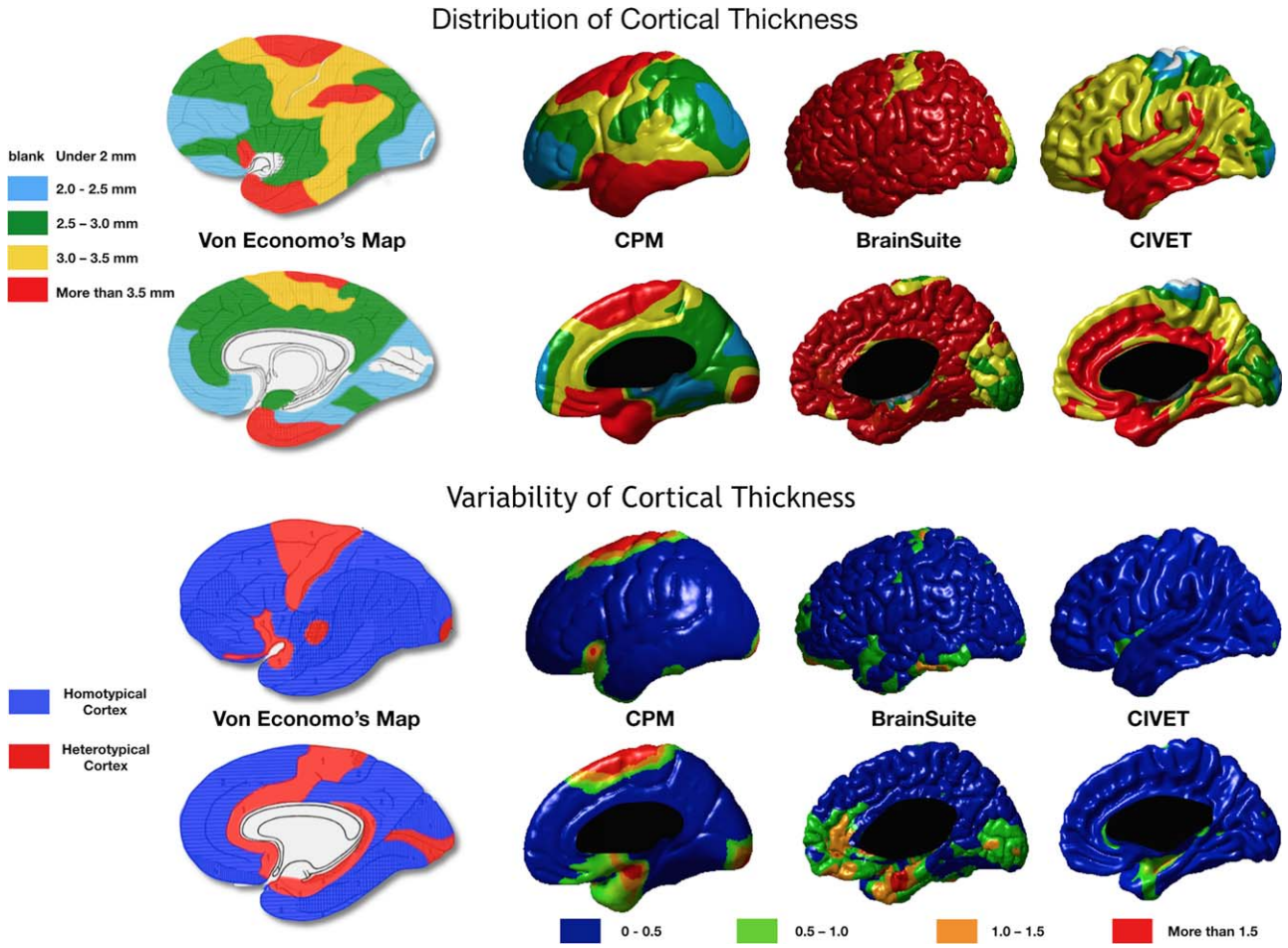

Figure 6.

CT distribution (Top) and variability (Bottom) maps derived from postmortem data (von Economo, 1929; Left) and neuroimaging data (outputs from the processing pipelines considered; Right). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

inspection suggests that, of the three pipelines, the data derived from the CPM pipeline might be most comparable to the von Economo's map. Further, consistent with previous histological findings (Kabani et al., 2001; Lerch, 2001), thickness data obtained by the CPM processing protocol showed more variability between individual brains in heterotypical than in homotypical regions. Nevertheless, note that recent high resolution 3D maps based on combined neuroimaging and histological data (such as the BigBrain project, Amunts et al., 2013), will be useful for obtaining a new "gold standard" for cross-validation studies.

Are Cognitively Matched Samples Equivalent in Their Brain Anatomical Configurations?

The straight answer is "no, they are not."

We applied an analytic strategy presumably appropriate for assessing inconsistencies related to the variability among samples. The complete group was divided into two subsamples carefully matched with respect to all cognitive domains, sex, age, and handedness. This was intended to maximize the likelihood of replication. Even when each SBM pipeline was consistent in estimating CT across cognitively matched subsamples (the patterns of distribution and variability were the same for the assessed subsamples for a given pipeline—see an example in Fig. 4), the brain correlates for psychologically matched samples failed to converge. These patterns were replicated on 100 additional pairs of matched subsamples.

Admittedly, we may lack statistical power. As suggested by Yarkoni (2009), an unpowered correlational analysis will consistently produce spatially circumscribed and numerically inflated effects. Thus, as the sample size grows, it is common to find progressively reduced correlation values. With the samples sizes used in the current study, only large effect sizes may be detected at an acceptable level of power. We would then expect that only a

fraction of the true effects in the population will be identified. Nonetheless, we might expect some level of replication for the fraction of effects that are identified.

An interesting finding is that most of the prominent results for the complete group were a combination of those obtained in both subsamples, irrespective of the SBM method. Thus, some effects become significant by increasing the sample size, whereas others become nonsignificant. The latter phenomenon occurs when the discovered effect in the complete group is mainly driven by one of the subsamples, or when the relationship between the psychological factor and CT is substantial for both subsamples, but in opposite directions. This let us to suggest that, with respect to the brain, intelligence, and other cognitive factors might be moving targets. The human brain is highly complex and there are huge individual differences (Mueller et al., 2013). Brains are general-purpose devices that may achieve equivalent cognitive goals using quite different neural networks. First, complex psychological outcomes are a composite of several simpler processes. Thus, an equivalent cognitive domain score (for instance in fluid intelligence) may be the result of specific cognitive configurations, varying across individuals in the efficiency with which each process involved is performed. The intersubjects variability in these specific configurations may have different neurobiological substrates.

Conversely, an alternative possibility is that there is no single neuroanatomical structure underlying a given higher order cognitive domain. This implies that there are different brain designs supporting the same function. Averaging large datasets of anatomical data (or functional signals) may help to hide the revealed inconsistencies, but this latter analytic strategy may also mask relevant information for a proper understanding of the complex brain-cognition relationship.

Concluding Remarks: What Is Next?

Here, we have emphasized the inhomogeneities across findings derived from surface-based structural neuroimaging studies. Our focus was on cognitive performance. Nevertheless, the lack of replicability found in the literature could be extended to other relevant psychological variables, particularly when complex behaviors are considered. Improved localization of "true" brain regions supporting a given function might be achieved by, for example, enlarging sample size, reporting results after correcting for multiple comparisons, or using enhanced neuroimaging techniques. These are methodological improvements, but the assumption that it is possible to locate unambiguous brain regions for a certain psychological outcome might still be flawed.

Meta-analyses demonstrated that there are no reliable function/structure-behavior associations (Yarkoni et al., 2010, 2011). Because many brain regions participate in apparently disparate functions, a selective correspondence is difficult to establish. Moreover, cognitive profiles are heterogeneous: people might display the same performance, but the way in which different underlying mental processes are involved may vary between them. Thus, we may expect to find a similar heterogeneity in the association between cognitive performance and brain morphology. A direct implication of this lack of consistency for structural brain properties—psychological functions relationships is the impossibility of assessing which processing protocol fits better with the expected results.

We acknowledge that large samples are needed. Pooling is becoming a popular strategy (e.g., Human Connectome Project; ADNI; ENIGMA), and our results suggest caution when applying this procedure. Statistical brute force might be insufficient and it can divert our attention from the main research goal, for instance, to find the brain substrate of individual differences in cognitive performance. Because of the dynamic nature of the human brain and the complexities of human cognition, replication needs carefully matched samples and strictly comparable psychological scores, neuroimaging methods, and brain properties. But even then replication of findings is not guaranteed, as suggested by the data presented and discussed here.

Imaging processing methods for functional and structural MR data are always under development to improve the biological plausibility of the findings. The present report was focused on past attempts, but new developments will be available soon. Thus, for instance, the new version of CIVET (2.0) is based on the fine-grained inputs provided by BigBrain, an ultrahigh-resolution 3D human brain model (Amunts et al., 2013). Similar improvements are observed in BrainSuite (v14a1; http://brainsuite.org/2014/06/brainsuite-14a-released/). It can be expected that the advances will help us to resolve the observed inconsistencies addressed in the current report. Therefore, we strongly recommend studies specifically addressing potential explanations of the instability between SBM pipelines outputs from a technical point of view. The pipelines used here comprise numerous image processing steps before running group analyses. All these steps vary widely across pipelines. Analyzing how each of these steps may influence the observed results requires specific methodological designs changing one step at a time and this challenge is far from the scope of the present study. Furthermore, external data obtained by other techniques are required for assessing the convergent validity of these morphological protocols.

Additionally, we encourage researchers to replicate their findings using different protocols in the same dataset, while clarifying each processing step and procedures used. As pointed out by Button et al. (2013), if the intended analyses produce null findings, they should be reported; if researchers decide to move on to explore the data in other ways to get significant findings, they must explicitly acknowledge this.

In summary, this study reveals a potential vulnerability in studies assessing the relationship between CT and behavior, which is in part attributable to methodological sources of instability. Solutions for the observed problems involve improvements in the algorithms needed for processing neuroimaging data from a surface-based approach

taking into account the information from other modalities for cross-validation.

One of our main challenges may be to rethink the way in which we are studying complex psychological factors at the biological level. Statistical analyses for identifying differentiable "brain profiles" may be useful for a better understanding of those brain networks properties supporting inter subject variability in cognitive performance.

REFERENCES

Ackerman PL, Beier ME, Boyle MO (2002): Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. J Exp Psychol Gen 131:567–589.

Abad FJ, Olea J, Ponsoda V, García C (2011): Medición en Ciencias Sociales y de la Salud. Madrid: Síntesis.

Ad-Dab'bagh Y, Lyttelton O, Muehlboeck JS, Lepage C, Einarson D, Mok K, Evans AC (2006): The CIVET image-processing environment: A fully automated comprehensive pipeline for anatomical neuroimaging research. In: Corbetta M, editor. Proceedings of the 12th Annual Meeting of the Organization for Human Brain Mapping. p S45. http://www.bic.mni.mcgill.ca/users/yaddab/Yasser-HBM2006-Poster.pdf

Amunts K, Lepage C, Borgeat L, Mohlberg H, Dickscheid T, Rousseau MÉ, Evans AC (2013): BigBrain: An ultrahigh-resolution 3D human brain model. Science 340:1472–1475.

Arbuckle JL (2007): AMOS, Version 16.0.1. Spring House, PA: Amos Development Corporation.

Benjamini Y, Hochberg Y. (1995): Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological):289–300.

Bennett GK, Seashore HG, Wesman AG (1990): Differential Aptitude Test, 5th ed. Madrid: TEA.

Bentler PM (1990): Comparative fit indexes in structural models. Psychol Bull 107:238–46.

Boucher M, Whitesides S, Evans A (2009): Depth potential function for folding pattern representation, registration and analysis. Med Image Anal 13:203–214.

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013): Power failure: Why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376.

Byrne BM (1998): Structural equation modelling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming. Erlbaum: Mahwah.

Cabeza R, Nyberg L (2000): Imaging cognition II: An empirical review of 275 PET and fMRI studies. J Cogn Neurosci 12:1–47.

Carroll JB, (1993): Human cognitive abilities. A survey of factor analytic studies. Cambridge: Cambridge University Press.

Carroll JB (2003): The higher-stratum structure of cognitive abilities: Current evidence supports g and about 10 broad factors. In H. Nyborg (Ed.), The scientific study of general intelligence: Tribute to Arthur R. Jensen (pp. 5–21). Amsterdam: Pergamon.

Cohen J (1988): Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Collins DL, Neelin P, Peters TM, Evans AC (1994): Automatic 3D intersubject registration of MR volumetric data in standardized talairach space. J Comput Assist Tomogr 18:192–205.

Colom R, Jung RE, Haier RJ (2007): General intelligence and memory span: Evidence for a common neuroanatomic framework. Cogn Neuropsychol 24:867–878.

Colom R, Abad FJ, Quiroga MA, Shih PC, Flores-Mendoza C (2008): Working memory and intelligence are highly related constructs, but why? Intelligence 36:584–606.

Colom R (2007): Intelligence? What intelligence?. Behavioral and Brain Sciences 30:155–156.

Colom R, Haier RJ, Head K, Álvarez-Linera J, Quiroga MA, Shih PC, Jung RE (2009): Gray matter correlates of fluid, crystallized, and spatial intelligence: Testing the P-FIT model. Intelligence 37(2): 124–135.

Colom R, Karama S, Jung RE, Haier RJ (2010a): Human intelligence and brain networks. Dialogues Clin Neurosci 12:489.

Colom R, Quiroga MA, Shih PC, Martínez K, Burgaleta M, Martínez-Molina A, Román FJ, Requena L Ramírez I (2010b): Improvement in working memory is not related to increased intelligence scores. Intelligence 38:497–505.

Cole MW, Yarkoni T, Repovš G, Anticevic A, Braver TS (2012): Global connectivity of prefrontal cortex predicts cognitive control and intelligence. J Neurosci 32:8988–8999.

Colom R, Burgaleta M, Román FJ, Karama S, Álvarez-Linera J, Abad FJ, Haier RJ (2013): Neuroanatomic overlap between intelligence and cognitive factors: Morphometry methods provide support for the key role of the frontal lobes. NeuroImage 72:143–152.

Deary I (2012): 125 years of intelligence in the American journal of psychology. Am J Psychol 125:145–154.

Fan X, Thompson B, Wang L (1999): Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. Struct Equ Modeling 6:56–83.

Frost MA, Goebel R (2012): Measuring structural–functional correspondence: Spatial variability of specialised brain regions after macro-anatomical alignment. Neuroimage 59:1369–1381.

Genovese CR, Lazar N, Nichols TE (2002): Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15:870–878.

Haier RJ, Colom R, Schroeder D, Condon C, Tang C, Eaves E, Head K (2009): Gray matter and intelligence factors: Is there a neuro-g? Intelligence 37:136–144.

Holmes CJ, Hoge R, Collins L, Woods R, Toga AW, Evans AC (1998): Enhancement of MR images using registration for signal averaging. J Comput Assist Tomogr 22:324–333.

Hu LT, Bentler PM (1999): Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Modeling 6:1–55.

Hunt EB (2011): Human Intelligence. Cambridge, UK: Cambridge University Press.

Johnson W, Bouchard T (2005): The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. Intelligence 33:393–416.

Jonides J, Lewis RL, Nee DE, Lustig CA, Berman MG, Moore KS (2008): The mind and brain of short-term memory. Annu Rev Psychol 59:193–224.

Jöreskog K (1993): Testing structural equation models. In: Bollen KA, Long JS Editors. Testing Structural Equation Models. Newbury Park: Sage. pp 294 − 315.

Joshi AA, Shattuck DW, Damasio H, Leahy RM (2012a): Geodesic curvature flow on surfaces for automatic sulcal delineation. In: 9th IEEE International Symposium on Biomedical Imaging (ISBI). IEEE. pp 430–433.

Joshi AA, Shattuck DW, Leahy RM (2012b): A method for automated cortical surface registration and labeling. In Biomedical Image Registration. Springer Berlin Heidelberg. pp 180–189.

Joshi SH, Cabeen RP, Joshi AA, Sun B, Dinov I, Narr KL, Woods RP (2012c): Diffeomorphic sulcal shape analysis on the cortex. IEEE Trans Med Imaging 31:1195–1212.

Jung RE, Haier RJ (2007): The Parieto-frontal integration theory (P-FIT) of intelligence: Converging neuroimaging evidence. Behav Brain Sci 30:135–154.

Kabani N, Le Goualher G, MacDonald D, Evans AC (2001): Measurement of cortical thickness using an automated 3-D algorithm: A validation study. Neuroimage 13:375–380.

Karama S, Colom R, Johnson W, Deary IJ, Haier R, Waber DP, Lepage C, Ganjavi H, Jung R, Evans AC, The brain development cooperative group (2011): Cortical thickness correlates of specific cognitive performance accounted for by the general factor of intelligence in healthy children aged 6 to 18. NeuroImage 55:1443–1453.

Kim JS, Singh V, Lee JK, Lerch J, Ad-Dab'bagh Y, MacDonald D, Evans AC (2005): Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a laplacian map and partial volume effect classification. Neuroimage 27:210–221.

Langer N, Pedroni A, Gianotti LR, Hänggi J, Knoch D, Jäncke L (2012): Functional brain network efficiency predicts intelligence. Hum Brain Mapp 33:1393–1406.

Lerch J (2001): Measuring Cortical Thickness. PhD thesis. McGill University Montreal.

Lerch JP, Evans AC (2005): Cortical thickness analysis examined through power analysis and a population simulation. Neuroimage 24:163–173.

Lyttelton O, Boucher M, Robbins S, Evans A (2007): An unbiased iterative group registration template for cortical surface analysis. Neuroimage 34:1535–1544.

MacDonald D (1998): A method for identifying geometrically simple surfaces from three dimensional images. PhD thesis, McGill University.

MacDonald D, Avis D, Evans AC (1994): Multiple surface identification and matching in magnetic resonance images. In: Visualization in Biomedical Computing. International Society for Optics and Photonics. pp 160–169.

MacDonald D, Kabani N, Avis D, Evans AC (2000): Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. NeuroImage 12:340–356.

McGrew K (2009): CHC theory and the human cognitive abilities project: standing on the shoulders of the giants of psychometric intelligence research. Intelligence 37:1–10.

Mazziotta J, Toga A, Evans A, Fox P, Lancaster J (1995): A probabilistic atlas of the human brain: Theory and rationale for its development. NeuroImage 2:89–101.

Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, Mazoyer B (2001): A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). Philos Trans R Soc Lond Series B Biol Sci 356:1293–1322.

Mueller S, Wang D, Fox MD, Yeo BT, Sepulcre J, Sabuncu MR, Liu H (2013): Individual Variability in Functional Connectivity Architecture of the Human Brain. Neuron 77: 586–595.

Naghavi HR, Nyberg L (2005): Common fronto-parietal activity in attention, memory, and consciousness: Shared demands on integration? Conscious Cogn 14:390–425.

Nisbett RE, Aronson J, Blair C, Dickens W, Flynn J, Halpern DF, Turkheimer E (2012): Intelligence: New findings and theoretical developments. Am Psychol 67:130–159. http://dx.doi.org/10.1037/a0026699.

Panizzon MS, Fennema-Notestine C, Eyler LT, Jernigan TL, Prom-Wormley E, Neale M, Kremen WS (2009): Distinct genetic influences on cortical surface area and cortical thickness. Cereb Cortex 19:2728–2735.

Raven J, Raven JC, Court JH, (2004): Manual for Raven's Progressive Matrices and Vocabulary Scales. San Antonio, TX: Harcourt Assessment.

Román FJ, Abad FJ, Escorial S, Burgaleta M, Martínez K, Álvarez-Linera J, Colom R (2014): Reversed hierarchy in the brain for general and specific cognitive abilities: A morphometric analysis. Hum Brain Mapp 35:3805–3818.

Sapiro G (2001): Geometric Partial Differential Equations and Image Analysis. Cambridge, UK: Cambridge UP.

Shattuck D, Sandor-Leahy S, Schaper K, Rottenberg DA, Leahy RM (2001): Magnetic resonance image tissue classification using a partial volume model. Neuroimage 13:856–876.

Sled JG, Zijdenbos AP, Evans AC (1998): A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 17:87–97.

Smith SM (2002): Fast robust automated brain extraction. Hum Brain Mapp 17:143–155.

Sowell ER, Thompson PM, Rex D, Kornsand D, Tessner KD, Jernigan TL, Toga AW (2002): Mapping sulcal pattern asymmetry and local cortical surface gray matter distribution in vivo: Maturation in perisylvian cortices. Cereb Cortex 12:17–26.

Sowell ER, Thompson PM, Leonard CM, Welcome SE, Kan E, Toga AW (2004): Longitudinal mapping of cortical thickness and brain growth in normal children. J Neurosci 24:8223–8231.

Talairach J, Tournoux P (1980): Co-planar Sterotaxic Atlas of the Human Brain. Stuttgart: Thieme.

Thompson PM, Hayashi KM, Sowell ER, Gogtay N, Giedd JN, Rapoport JL, Toga AW (2004): Mapping cortical change in alzheimer's disease, brain development, and schizophrenia. Neuroimage 23:S2–S18.

Thurstone L (1938): Primary Mental Abilities. Psychometric Monographs. 1 University of Chicago Press.

Tohka J, Zijdenbos A, Evans A (2004): Fast and robust parameter estimation for statistical partial volume models in brain MRI. Neuroimage 23:84–97.

von Economo CF (1929): The Cytoarchitectonics of the Human Cerebral Cortex. Humphrey Milford University Press.

Winkler AM, Kochunov P, Blangero J, Almasy L, Zilles K, Fox PT, Glahn DC (2010): Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. Neuroimage 53:1135–1146.

Worsley KJ, Taylor JE, Tomaiuolo F, Lerch J (2004): Unified univariate and multivariate random field theory. Neuroimage 23(Suppl 1):S189–195. −

Yarkoni T (2009): Big correlations in little studies: Inflated fMRI correlations reflect low statistical power-Commentary on Vul et al. Perspectives on Psychological Science 4(3): 294–298.

Yarkoni T, Poldrack RA, Van Essen DC, Wager TD (2010): Cognitive neuroscience 2.0: building a cumulative science of human brain function. Trends in cognitive sciences 14: 489–496.

Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD (2011): Large-scale automated synthesis of human functional neuroimaging data. Nature methods 8:665–670.

Yela M (1969): Rotación de Figuras Macizas (Rotation of solid figures). Madrid: TEA.

Zhao L, Boucher M, Rosa-Neto P, Evans AC (2012): Impact of scale space search on age-and gender-related changes in MRIbased cortical morphometry. Human Brain Mapping 34:2113–2138.

Zijdenbos AP, Forghani R, Evans AC (2002): Automatic "pipeline" analysis of 3-D MRI data for clinical trials: Application to multiple sclerosis. IEEE Trans Med Imaging 21:1280–1291.