Neuroscience 297 (2015) 38–46

BEHAVIORAL AND fMRI EVIDENCE OF THE DIFFERING COGNITIVE LOAD OF DOMAIN-SPECIFIC ASSESSMENTS

S. J. HOWARD,[a]* H. BURIANOVÁ,[b] J. EHRICH,[c] L. KERVIN,[a] A. CALLEIA,[a] E. BARKUS,[d] J. CARMODY[e] AND S. HUMPHRY[f]

[a] *School of Education, University of Wollongong, New South Wales 2522, Australia*

[b] *Centre for Advanced Imaging, University of Queensland, Queensland 4072, Australia*

[c] *Faculty of Education, Monash University, Victoria 3800, Australia*

[d] *School of Psychology, University of Wollongong, New South Wales 2522, Australia*

[e] *Neurology Department, Wollongong Hospital, New South Wales 2500, Australia*

[f] *Graduate School of Education, University of Western Australia, Western Australia 6009, Australia*

Abstract—Standards-referenced educational reform has increased the prevalence of standardized testing; however, whether these tests accurately measure students' competencies has been questioned. This may be due to domain-specific assessments placing a differing domain-general cognitive load on test-takers. To investigate this possibility, functional magnetic resonance imaging (fMRI) was used to identify and quantify the neural correlates of performance on current, international standardized methods of spelling assessment. Out-of-scanner testing was used to further examine differences in assessment results. Results provide converging evidence that: (a) the spelling assessments differed in the cognitive load placed on test-takers; (b) performance decreased with increasing cognitive load of the assessment; and (c) brain regions associated with working memory were more highly activated during performance of assessments that were higher in cognitive load. These findings suggest that assessment design should optimize the cognitive load placed on test-takers, to ensure students' results are an accurate reflection of their true levels of competency.

Key words: cognitive load, standardized testing, assessment, spelling, fMRI, educational neuroscience.

INTRODUCTION

A fundamental aim of educational assessment is to maximize validity and reliability in measuring students' abilities (Borsboom et al., 2004). In pursuit of this aim, standards-based educational reform has increased the prevalence of standardized testing and the stakes associated with students' results on these tests (Pellegrino, 2001). In fact, internationally, schools are often funded and publicly ranked based on these results. Yet, the extent to which these tests accurately index students' competencies has been questioned (Pellegrino, 2001; Wiliam, 2003). Specifically, it has been argued that many standardized national curriculum assessments may also assess domain-general (i.e., general purpose, content-free) cognitive capacities in the attempt to assess literacy and numeracy knowledge and skills (Willet and Gardiner, 2009). In support of this suggestion, neuroimaging research suggests that even the simplest literacy and numeracy tasks engage domain-general cognitive networks (Baddeley, 2003; Knudsen, 2007). The domain-general resource most commonly implicated in students' performance on standardized assessments is working memory, whose capacity-limited nature constrains the amount of information that concurrently can be activated, maintained, and manipulated in mind (Engle, 2010). It is therefore unclear whether standardized assessment results reflect students' true literacy and numeracy competencies or whether their scores have been restricted by the limits of their domain-general cognitive resources (e.g., the cognitive demands of the assessment outpacing students' available working memory capacity).

The effects of divergent domain-general cognitive demands are evidenced by research indicating that children's ability to demonstrate their knowledge and skills varies by type of assessment. For instance, in the area of literacy assessment, a recent study found that 75% of students were better able to spell dictated words than correct visually presented misspelled words (the latter based on Australia's National Assessment Program – Literacy and Numeracy, or NAPLAN, method of spelling assessment; Willet and Gardiner, 2009). This finding is consistent with additional studies suggesting that error correction and proofreading tasks typically involve more than just spelling ability (Croft, 1982; Frisbie and Cantor, 1995; although for conflicting results,

*Corresponding author. Tel: +61-2-4221-5165.
E-mail addresses: stevenh@uow.edu.au (S. J. Howard), hana.burianova@cai.uq.edu.au (H. Burianová), john.ehrich@monash.edu.au (J. Ehrich), lkervin@uow.edu.au (L. Kervin), amc998@uowmail.edu.au (A. Calleia), ebarkus@uow.edu.au (E. Barkus), john.carmody@sesiahs.health.nsw.gov.au (J. Carmody), stephen.humphry@uwa.edu.au (S. Humphry).

Abbreviations: BOLD, blood oxygenation level-dependent; fMRI, functional magnetic resonance imaging; LV, latent variables; MI, misspelled identified; MNI, Montreal Neurological Institute; MU, misspelled unidentified; NAPLAN, National Assessment Program – Literacy and Numeracy; PLS, Partial Least Squares.

http://dx.doi.org/10.1016/j.neuroscience.2015.03.047
0306-4522/

see Westwood, 1999). This suggests at least some variability in spelling performance may be related to individual differences in domain-general cognitive abilities. Specifically, correcting misspelled words may also require the cognitive flexibility to switch between orthographic representations, thereby placing greater demands on working memory. In fact, working memory has been shown to underlie performance on a broad range of standardized and educational assessments (Gathercole et al., 2003; Strattman and Hodson, 2005; Alloway and Gregory, 2012) and is a particularly powerful predictor of academic achievement (including literacy and numeracy achievement; Blair and Razza, 2007; Best et al., 2011).

Cognitive load researchers have similarly highlighted how the complexity of information and its method of presentation can overwhelm children's limited working memory capacity (van Merrienboer and Sweller, 2005; Kirschner et al., 2011), thus restricting students' ability to acquire and demonstrate their emerging academic competencies. Although fundamentally a theory of learning and instructional design, Cognitive Load Theory principles are similarly applicable to educational assessment in that assessment, like instruction, can impose more or less demand (cognitive load) on test-takers' working memory. Differences in cognitive load across assessments can occur as a function of the inherent complexity of the knowledge and skills being assessed (intrinsic load), immaterial aspects of the assessment relative to the knowledge and skills being assessed (extraneous load), and the mental effort expended on assessment-relevant processes (germane load). For instance, the assessment of whether a student can spell a particular word can be described as being low in element interactivity (successful performance requires minimal reference to, or interaction of, other learned concepts or procedures; Sweller, 1994) compared to correcting a misspelling of that same word. The latter imposes a higher cognitive load, although the specific type of load imposed is less clear. That is, if the assessment aimed to evaluate students' proofreading abilities, the additional load could be characterized as intrinsic load (although this would be an assessment of, at least partly, different knowledge and skills than spelling). However, if the assessment aimed to measure the level of complexity at which students could accurately spell, the additional load could be characterized as extraneous (in that proofreading is a non-essential process for producing the correct spelling of a word). More than just semantics, it is notable that many large-scale, national assessment programs characterize the knowledge and skills they assess using identical terms (e.g., 'spelling'), yet assess these abilities in a highly disparate manner. As a consequence, these assessments may vary in the cognitive demands placed on test-takers' working memory, even when the domain-specific knowledge and skills they assess remain constant. This has important implications for interpretation of assessment results (especially given individual differences in working memory capacity and the resulting differential effect on test performance that may occur) and designing appropriate educational experiences to foster the assessed knowledge and skills.

Although this issue of the domain-general demands of domain-specific assessments is derived from education, it is not easily addressed by traditional educational research methods. For example, neither qualitative nor behavioral studies of spelling assessment are able to conclusively determine the extent to which observed performance differences are spurious (e.g., due to situational or motivational factors), transitory (e.g., due to temporary practice effects), or the product of more fundamental cognitive processes underlying learning and performance (e.g., the varied cognitive load of different modes of assessment). This is an ongoing issue for educational psychologists. Mechanisms of learning and performance are too often defined in operationist terms as psychometric constructs measured exclusively by tests (Michell, 2005; Kelly, 2011), which often are not founded upon substantive theory or an understanding of the function of the brain. The emerging field of educational neuroscience, in contrast, seeks to leverage insights from education, psychology and neuroscience to bridge the gap between the conscious mind and living brain (Szucs and Goswami, 2007). One advantage of applying neuroscientific methods to educational issues is that the contributions of individual neural systems to academic achievement (including domain-general systems) can be identified and quantified (Vander Wyk and Pelphrey, 2011). These neuroanatomical findings can reconcile emerging brain-based insights (such as brain-based evidence of the cognitive load of different forms of assessment) with established educational theory (such as Cognitive Load Theory) to support, refine or advance long-regarded principles of educational best practice (e.g., Whelan, 2007).

The current study sought to combine neuroscientific and behavioral research methods to examine the extent to which domain-general neural correlates contribute to performance on different modes of assessment. Specifically, functional magnetic resonance imaging (fMRI) was used to identify and quantify the domain-general contributions facilitating performance on three different spelling assessments (adapted from Australia's NAPLAN tests, the UK's National Curriculum Tests, and commercial standardized spelling assessments). In addition, out-of-scanner spelling assessments were used to further investigate the relationship between brain (i.e., domain-general neural networks) and behavior (i.e., assessment results). It was expected that triangulation of these results would provide neurological and behavioral evidence that spelling assessments differ in the cognitive load they place on test-takers, as evidenced by: (a) decreased spelling performance on assessments that are higher in cognitive load; and (b) working memory accounting for important variance on assessments that impose greater cognitive load. Specifically, it was expected that error correction methods of spelling assessment (i.e., identify and correct a misspelled word, in line with NAPLAN's method of spelling assessment) would impose greater cognitive load on test takers than dictation forms of assessment (i.e., spell the dictated word, in line with the UK's National Curriculum Tests). As a consequence of

this predicted difference in cognitive load, error correction assessments were expected to additionally recruit areas of the frontoparietal network, which are associated with working memory and increased attention (e.g., prefrontal and parietal cortices; Corbetta and Shulman, 2002; Ashby et al., 2005).

Although spelling is considered by some to be a 'constrained skill' (Paris, 2005), it was adopted here to investigate the cognitive load of different forms of assessment due to: (i) the compatibility of spelling assessments with fMRI restrictions; (ii) the consistency with which spelling is assessed through standardized assessments internationally; and (iii) the established link between spelling and reading, insofar as both are found to rely on similar knowledge and skills (Westwood, 2008) and improved spelling has been suggested to enhance subsequent reading ability (Graham et al., 2002; Santoro et al., 2006). It therefore follows that, although the current study focused on spelling assessment, the insights generated can inform principles of assessment development and (re)design more broadly.

EXPERIMENTAL PROCEDURES

Participants

Participants were 12 university students recruited from two large Australian universities. All participants were healthy adults aged 18–35 years (M = 22.00, SD = 2.20; range = 19.11–26.50 years) with normal or corrected-to-normal vision and no prior history of neurological or psychological impairment. Two-thirds of participants were female (n = 8). All participants were native speakers of English. Participants gave written informed consent as a requirement for participation. This research was conducted in accordance with the Declaration of Helsinki and approval to conduct this research was obtained from the participating institutions' human subjects review boards.

Measures

Spelling proficiency was assessed under each of the following four experimental conditions (ordered from highest to lowest cognitive load, in line with our hypotheses): (1) auditory and textual presentation of a sentence with a *misspelled unidentified* (MU) word to be identified and then corrected (e.g., 'Doctors innoculate their patients to prevent illnesses such as smallpox'); (2) auditory and textual presentation of a sentence with a *misspelled identified* (MI) word to be corrected (e.g., 'The shops are in close *proksimity* to my house'); (3) auditory and textual presentation of a sentence with a missing word (*Blank*) for spelling (e.g., 'My painting is _____ compared to that masterpiece'); and a *control* condition involving (4) auditory and textual presentation of a sentence with an identified, correctly spelled word to be spelled (e.g., 'The chocolate looked *irresistible*'). The first two conditions were based on Australia's NAPLAN tests and the third condition was based on the UK's National Curriculum Tests. Words to be spelled for all conditions had been identified as age-appropriate by standardized adult literacy assessments. Sentences were developed for each word and the resultant items were piloted (N = 31). Sentences for the current study were selected on the basis of accuracy between 10% and 95% in this pilot. This criterion yielded 120 sentences, which were divided evenly into the four conditions. Each condition was then administered twice per fMRI scan (i.e., 15 sentences per run). Sentence stimuli for each run and condition were balanced on the basis of pilot test accuracy (M = 0.64, SD = 0.25; Calleia and Howard, 2014), word frequency norms ($M_{freq/500}$ = 9.18, SD = 11.83; Francis and Kucera, 1982), word length ($M_{\#letters}$ = 9.11, SD = 2.10), and type of misspelling (e.g., omission of a letter, substitution of a letter, addition of a letter, homophone; all ps < .05).

Procedure

Participants completed a 10-min familiarization training within 24 h of their scan, in which they were provided with demonstration of the test types and in-scanner requirements (e.g., button press protocols). During scanning, participants completed eight 15-word spelling tests (divided into runs, such that each experimental condition was presented twice) over the course of a 90-min scan. These eight runs were presented in pseudo-random order (i.e., the experimental conditions were presented in random order, but no condition was repeated until each was presented once). Runs proceeded as follows: (1) instructions for 30 s, which stated condition requirements; (2) fixation for 4 s; (3) auditory and textual presentation of a sentence for 30 s; and (4) repetition of (2) and (3) for another 14 sentences. Participants responded by: (a) listening to and reading the sentence; (b) mentally identifying the word to be spelled and its proper spelling (search phase); (c) pressing and holding a button to indicate the beginning of the spelling phase (during which participants covertly spelled the target word); and (d) releasing the button to indicate completion of the spelling phase (see Fig. 1). Participants automated this process (supported by a >98% rate of protocol compliance in scanner) in the pre-scan training. Immediately post-scan, participants were administered a written spelling test identical to those presented in-scanner, on which participants identified how they spelled each word in the scanner.

fMRI data acquisition

Anatomical and functional images were acquired at the University of Sydney's Brain and Mind Research Institute in Sydney, Australia, using a GE Discovery MR750 whole body 3T scanner with a matrix 8-channel head coil. Anatomical images were acquired using 196 axial slices, TR = 7.21 s, TE = 2.76 s, flip angle = 12°, TI = 450, voxel size = 0.9 mm^3, acquisition matrix = 256 × 256. Brain activation was assessed using the blood oxygenation level-dependent (BOLD) effect (Ogawa et al., 1990) with optimal contrast. Functional images were obtained using 45 axial slices, TR = 3 s, TE = 30 ms, flip angle = 90°, FOV = 240 mm, voxel

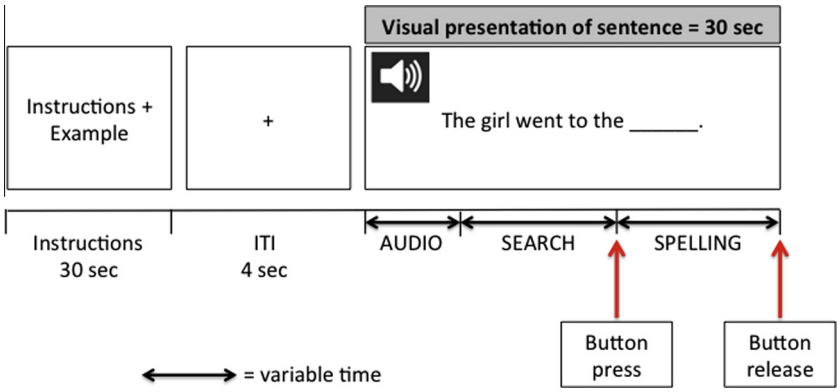

Fig. 1. In-scanner experimental paradigm. At the beginning of each run, instructions for the upcoming condition and an example sentence were presented for 30 s, followed by a 4-s inter-trial-interval (ITI), and then a 30-s trial during which a sentence was presented audio-visually. The visual sentence remained on the screen for the entire 30 s, whereas auditory presentation varied with the length of the sentence (offset of which signaled the start of the search phase). When ready to spell, participants pressed the button (signaling the end of the search phase and start of the spelling phase), releasing it once spelling was completed (signaling the end of the spelling phase).

size = 1.875 × 1.875 × 3 mm, acquisition matrix = 128 × 128.

fMRI data preprocessing and analysis

The acquired fMRI images were preprocessed using the Statistical Parametric Mapping software (SPM8; http://www.fil.ion.ucl.ac.uk/spm). First, functional images were slice-timing corrected and then realigned onto the mean image for head-motion correction. The anatomical image was then segmented and spatially normalized to the T1-weighted Montreal Neurological Institute (MNI) template, and the normalization parameters were applied to the functional data. Finally, data were spatially smoothed by convolving each volume with an isotropic Gaussian kernel (FWHM = 6 mm). For the analysis, all trials of each condition were averaged within and across the condition's two runs.

The fMRI data were analyzed using Partial Least Squares (PLS) analysis (McIntosh et al., 1996; Krishnan et al., 2011). PLS is a multivariate technique that examines covariance between activity in all brain voxels and experimental conditions, providing sets of mutually independent spatial patterns depicting brain regions that show the strongest relationship to the contrasts across conditions. Using PLS, latent variables (LV), defined as cohesive patterns of neural activity associated with a task, were identified (the LV accounting for the greatest covariance is extracted first) across conditions. Of primary interest was brain activity during the search phase, for which distinct patterns of activation were expected across experimental conditions due to differing processes required to plan a response (whereas the spelling phase involved identical processes across experimental conditions). We therefore isolated activity during the search phase (starting at the offset of auditory presentation of the sentence and ending at onset of spelling, as indicated by a button press) and spelling phase (starting at button press and ending at button release) as distinct events for our event-related analyses. Activity at each time point in the analysis was normalized to activity in the onset timepoint. Our measure of each phase-related activity was thus relatively uninfluenced by activity in the rest of the trial.

A permutation test determined significance of each LV and a bootstrap estimation of the standard errors determined the reliability of each LV (Efron, 1985). Peak voxels with a salience/SE ratio > 3.0 were considered to be reliable, as this approximates $p < .003$ (Sampson et al., 1989).

RESULTS

Data screening

The out-of-scanner spelling test data was first explored to identify any invalid trials in the scanner (those in which participants indicated that they either misheard or misinterpreted the target word). This resulted in a loss of 7.2% of the data. Subsequent analyses considered only remaining (valid) data. Rasch analysis of spelling data was also conducted to evaluate the psychometric properties of the spelling tests. General rules of thumb for a Rasch analysis of dichotomous variables require around 10 persons per item to ensure meaningful analysis (Andrich et al., 2005). However, due to the small number of persons ($N = 12$) relative to a larger number of items ($n = 90$), the data matrix was transposed so the variables associated with items were analyzed as persons, and vice versa. The symmetry of person and item parameters in the model permits such a transposition. Rasch analysis of these data revealed a non-significant item-trait interaction, $X^2 = 32.62$, $p = .112$, indicating good overall fit of the data to the Rasch model. The PSI – an index of internal consistency similar to Cronbach's alpha – of .85 indicated good reliability of the test. Taken together these results suggested a valid and reliable scale.

Behavioral spelling performance

To evaluate the effect of experimental condition on participants' spelling accuracy, a repeated-measures ANOVA was run on the proportional accuracy scores for

each condition. Results indicated a main effect of Condition, $F(2,22) = 7.33$, $p = .004$, $\eta^2 = .40$. Post-hoc analyses indicated that accuracy was highest in the Blank condition ($M = 0.74$, $SD = 0.20$), followed by the misspelled identified condition ($M = 0.65$, $SD = 0.17$) and the Misspelled Unidentified conditions ($M = 0.63$, $SD = 0.17$), which did not significantly differ. These results were consistent with our hypotheses, insofar as performance was highest in the low cognitive load condition (Blank) compared to conditions predicted to be higher in cognitive load (MI, MU). This result was subsequently explored in relation to the fMRI data.

fMRI results

To assess the neural correlates of the experimental conditions, PLS analyses were carried out comparing brain activation during the search and spelling phases of each condition. During the search phase, two significant patterns of large-scale activity were identified. The first pattern differentiated the Blank condition from both the misspelled identified (MI) and misspelled unidentified (MU) conditions, accounting for 69% of covariance in the data (see Table 1 and Fig. 2). A large-scale distributed network showed higher activation in the MI and MU conditions than in the Blank condition, including bilateral frontoparietal network, temporal regions, basal ganglia, and caudate nucleus (Fig. 2a), whereas the Blank condition activated bilateral angular gyrus, posterior cingulate gyrus, right middle and medial frontal cortices, parahippocampus, and occipital cortices (Fig. 2b). In contrast to the Blank condition, which engaged areas important for semantic processing, concept retrieval, and conceptual integration (Binder et al., 2009), the MI and MU conditions indicated a greater, likely intrinsic (Whelan, 2007), cognitive load by engaging nodes of the dorsal attentional and working memory networks (Corbetta and Shulman, 2002; Ashby et al., 2005).

The second identified pattern differentiated the MI from the MU condition, accounting for 31% of covariance in the data (see Table 2 and Fig. 3). In contrast to MI, MU engaged bilateral inferior parietal lobule and fusiform gyrus, left dorsolateral and inferior prefrontal cortices, left hippocampus, and putamen (Fig. 3a), reflecting the engagement of areas that have

Table 1. Regions differentially engaged during Blank and MI/MU experimental conditions

Region	Hem	BA	MNI coordinates			Ratio
			x	*y*	*z*	
MI and MU > Blank						
Dorsolateral prefrontal cortex	L	9	−38	8	34	8.12
	R	9	52	20	28	3.92
Inferior parietal lobule	L	40	−44	−38	42	4.23
	R	40	42	−38	42	6.14
Superior parietal lobule	L	7	−24	−64	58	6.85
	R	7	22	−58	62	5.92
Fusiform gyrus	L	19	−10	−70	−6	4.51
	R	19	16	−72	−6	5.44
Putamen	L		−22	4	−6	5.78
	R		20	6	−10	9.55
Middle temporal gyrus	L	21	−38	−80	22	7.48
	R	21	52	−76	20	4.25
Superior temporal gyrus	L	22	−50	−48	16	4.76
	R	22	56	−40	16	3.24
Middle occipital gyrus	L	19	−44	−80	4	4.92
	R	19	46	−84	4	9.05
Blank > MI and MU						
Lingual gyrus	L	18	−6	−78	−8	4.83
	R	18	16	−72	−6	5.44
Parahippocampus	L	19	−36	−44	2	5.23
Hippocampus	R		26	−44	0	4.50
Medial frontal gyrus		10	2	62	0	4.43
Thalamus	R		20	−10	16	3.73
Caudate nucleus	L		−20	18	18	4.26
Cuneus	L	18	−14	−92	18	5.18
	R	18	16	−90	18	4.63
Posterior cingulate gyrus		31	4	−48	34	7.17
Angular gyrus	L	39	−42	−66	36	3.42
	R	39	44	−64	38	3.23
Middle frontal gyrus	R	8	30	36	42	4.27

Note: Hem = hemisphere; R = right; L = left; BA = Brodmann's area; Ratio = salience/SE ratio from the bootstrap analysis; *x* coordinate = right/left; *y* coordinate = anterior/posterior; *z* coordinate = superior/inferior.

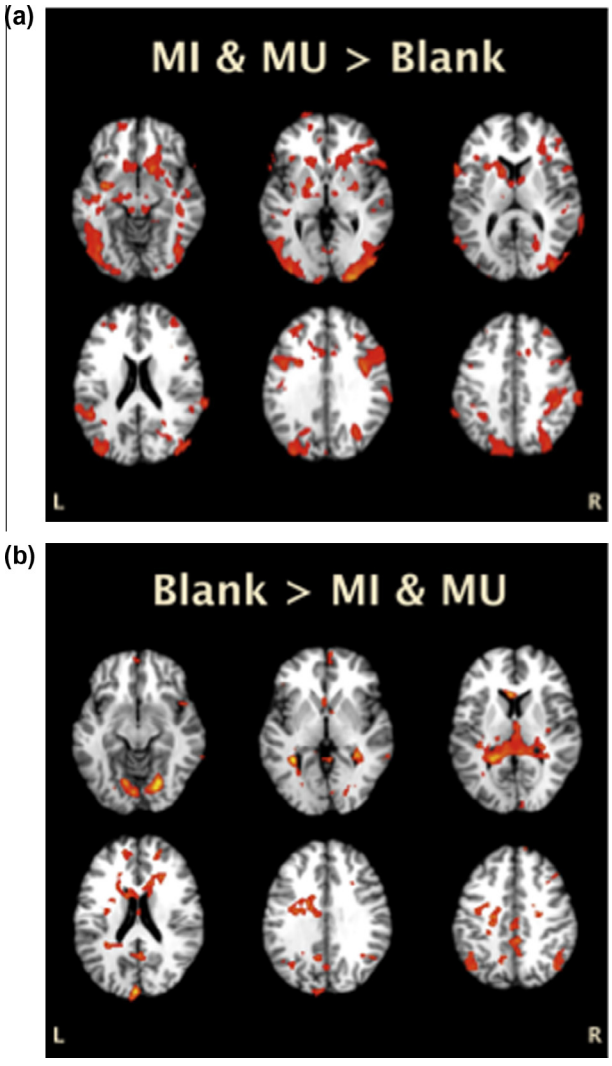

Fig. 2. (a) Axial slices illustrating a pattern of whole-brain activity during the mental search phase of the MI and MU conditions relative to Blank and (b) a pattern of whole-brain activity during the search phase of the Blank condition relative to MI and MU conditions. L = left hemisphere; R = right hemisphere.

been shown to be active during the retrieval of target concepts in a contextually weak semantic environment (Zempleni et al., 2007). In contrast, MI engaged bilateral angular gyrus, insula, left caudate nucleus, and medial prefrontal cortices, reflecting semantic processing and automatic comprehension of the identified, to-be-spelled word (Binder et al., 2009; Fig. 3b). During the spelling phase, all spelling conditions activated a whole-brain pattern, which included the anterior and posterior cingulate gyrus, angular and supramarginal gyri, superior frontal gyrus, insula, and parahippocampus, accounting for 71% of covariance in the data (see Table 3 and Fig. 4).

DISCUSSION

This study sought to examine the domain-general (working memory) cognitive demands of current domain-specific educational assessments. Research in this area is important given the increased prevalence of standardized

Table 2. Regions differentially engaged during MI and MU experimental conditions

Region	Hem	BA	MNI coordinates			Ratio
			x	*y*	*z*	
MU > MI						
Inferior parietal lobule	L	40	−36	−40	42	4.28
	R	40	44	−38	46	5.12
Inferior frontal gyrus	L	9	−32	28	−2	3.92
Dorsolateral prefrontal cortex	L	9	−48	10	28	3.70
Putamen	L		−16	6	−6	5.49
	R		18	10	−4	4.01
Hippocampus	L		−26	−40	−4	6.98
Fusiform gyrus	L	37	−42	−50	−16	8.71
	R	37	40	−58	−10	5.21
MI > MU						
Angular gyrus	L	39	−46	−48	38	4.48
	R	39	48	−62	38	4.78
Caudate nucleus	L		−22	−4	26	5.35
Posterior insula	L	13	−42	−34	20	4.93
	R	13	54	−28	22	4.05
Medial frontal gyrus		10	−16	54	−2	3.21
		8	6	54	34	7.30

Note: Hem = hemisphere; R = right; L = left; BA = Brodmann's area; Ratio = salience/SE ratio from the bootstrap analysis; *x* coordinate = right/left; *y* coordinate = anterior/posterior; *z* coordinate = superior/inferior.

educational testing around the world and the high stakes associated with students' results on these tests. Our data provide converging evidence that domain-specific spelling assessments, based upon current international and commercial methods of assessment, differ in the cognitive load that they place on test-takers. Specifically, error correction methods of spelling assessment (MI and MU), which were hypothesized to impose greater cognitive load on test-takers, displayed increased recruitment of neural areas associated with working memory and decreased performance compared to the production of correct spellings (Blank). Thus, claims that different methods of domain-specific educational assessment index students' competencies in a consistent manner appear questionable.

Specifically, consistent with prior studies (Croft, 1982; Frisbie and Cantor, 1995; Willet and Gardiner, 2009), our behavioral results indicated that participants performed better in the production condition (based upon the UK's National Curriculum Tests) than in the error correction conditions (based upon Australia's NAPLAN tests). This was the case despite the tests being balanced on the basis of word length, difficulty, and frequency. This finding is consistent with suggestions that error correction (proofreading) forms of spelling assessment may require additional domain-general processes to overcome interference from plausible (but incorrect) letter sequences, to activate correct orthographic representations. Although other studies have found significant correlations between error correction, proofreading, production, and multiple-choice forms of spelling assessment (with correlations ranging from .77 to .97), it is noted that these tests often involved highly discrepant task requirements (e.g., proofreading tasks requiring

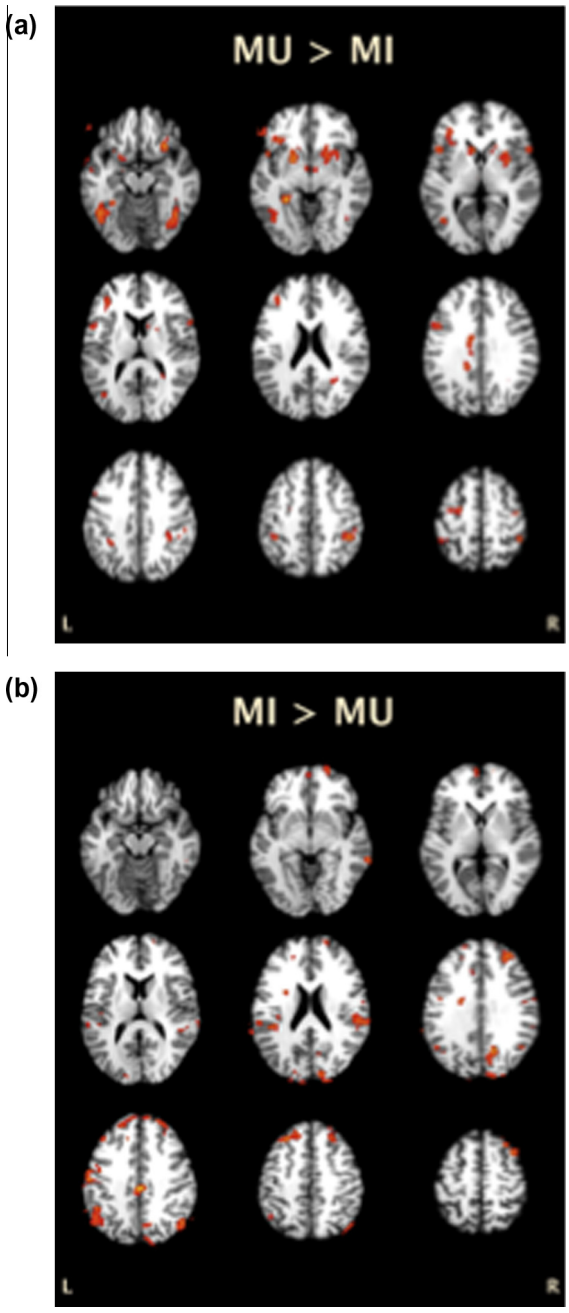

Fig. 3. (a) Axial slices illustrating a pattern of whole-brain activity during the mental search phase of MU relative to MI spelling condition and (b) a pattern of whole-brain activity during the search phase of MI relative to MU. L = left hemisphere; R = right hemisphere.

identification of misspelled words without students correcting them; Westwood, 1999), consequently complicating interpretation of these results.

A potential explanation for discrepant performance across equivalent spelling tests is that different modes of assessment may differ in the cognitive load they place on test-takers. That is, Cognitive Load Theory suggests that information varies in the demands (cognitive load) it places on learners' working memory, as a function of inherent complexity of the information (intrinsic load) and the complexity with which the

Table 3. Regions differentially engaged during spelling and fixation

Region	Hem	BA	MNI coordinates			Ratio
			x	*y*	*z*	
Spelling > Fixation						
Anterior cingulate gyrus		24	0	38	0	6.44
Posterior cingulate gyrus		31	4	−62	32	12.94
Superior frontal gyrus	R	8	24	26	48	10.06
Angular gyrus	L	39	−46	−60	38	5.27
	R	39	46	−64	40	9.48
Supramarginal gyrus	L	40	−54	−54	34	3.43
	R	40	56	−50	32	7.23
Precuneus		7	4	−58	44	6.55
Insula	L	13	−34	−20	20	4.93
	R	13	36	−24	22	10.65
Parahippocampus	L	36	−30	−48	−4	9.16
	R	36	30	−40	−4	6.23
Lingual gyrus	R	18	32	−68	−4	4.30

Note: Hem = hemisphere; R = right; L = left; BA = Brodmann's area; Ratio = salience/SE ratio from the bootstrap analysis; *x* coordinate = right/left; *y* coordinate = anterior/posterior; *z* coordinate = superior/inferior.

information is presented (extraneous load; van Merrienboer and Sweller, 2005; Kirschner et al., 2011). Although the foremost concern of Cognitive Load Theory is designing instruction and learning experiences that are founded upon a knowledge of human cognitive architecture, our results suggest that these Cognitive Load principles may similarly apply to assessment of student aptitudes (rather than applying solely to the acquisition of these competencies). For instance, it has been suggested that the two-step process in error correction and proofreading tests (proofreading, then correction) may require more, or more complex, activation and manipulation of information in working memory compared to production of correct spellings (Pearson, 2012). Although this assertion was made without empirical support, brain-based evidence for this suggestion is derived from our finding that a frontoparietal network, often associated with working memory (Corbetta and Shulman, 2002; Ashby et al., 2005), was more highly activated in the error correction conditions compared to the production condition.

Whelan (2007) suggested mapping of fMRI activations to specific sources of cognitive load (which provides a potentially more valid and reliable means to measure cognitive load than the existing dual-task, self-report, or physiological methods) suggests that the additional load in the error correction conditions may have been the result of its increased intrinsic load. That is, the processes involved in error correction may be inherently more complex (higher in element interactivity) than the production of a correct spelling. In the current study, this was supported by increased activation during error correction in the dorsolateral prefrontal cortex, which Whelan (2007) aligns with intrinsic load (whereas germane load is aligned to networks underlying motivation and extraneous load is aligned with the modulation of attention across sensory modalities).

Although comparisons and rankings of individuals, schools, and geographic regions are typically conducted

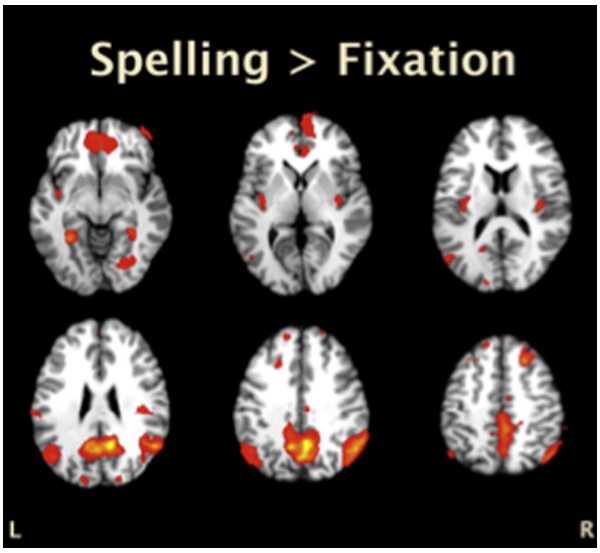

Fig. 4. Axial slices illustrating a pattern of whole-brain activity during the mental search phase of all spelling conditions vs. fixation. L = left hemisphere; R = right hemisphere.

within assessments, this disparity in cognitive load across assessments is nevertheless problematic insofar as large-scale standardized educational assessment programs often conceptualize the knowledge and skills they assess in a similar manner (e.g., 'spelling'), despite differences in their methods of assessment. As such, interpretation of an assessment's results and attempts to foster students' associated knowledge and skills are complicated by the lack of clarity regarding what assessments actually measure. The current study highlights the importance of assessment being clearly aligned with, and derived from, the intended learning outcomes. For instance, assessments aiming to evaluate students' ability to produce correct spellings, yet requiring the students to correct spelling errors, may not yield results reflecting students' true competencies in this area (increased extraneous load overwhelming working memory). In contrast, if an assessment aims to assess students' ability to locate and correct errors, clear statement of this aim would allow educators, parents, and students to align their teaching, learning, and efforts with these foundational abilities. While this study focused on standardized methods of assessment, the imposition of unnecessary, extraneous load is also an important consideration for educators as they develop their own classroom assessments.

Although it can be argued that our method of spelling assessment itself carried additional cognitive load, in that it required participants to mentally coordinate their response instead of transcribing their answers, it is critical to note that this requirement remained constant across conditions. As such, performance on error correction conditions involved additional domain-general resources over and above those associated with the response method. Nevertheless, given the constraints of fMRI data collection, which prevented more traditional methods of responding (e.g., no speech or head movement), future research using alternative neural recording methods is required to explore the cognitive load of educational assessments in more traditional testing contexts. Electroencephalography, for instance, has been suggested as one possible means by which to estimate cognitive load (Murata, 2005). Extending this investigation to the assessment of children will also be important to examine whether the same patterns of performance and neural activations are evident across a range of ages and expertise. Yet assessment is not exclusive to young children. It is thus expected that the assessment principles derived from this study can inform assessment design more broadly, including at the secondary and tertiary levels.

CONCLUSION

This study provides converging behavioral and neural evidence that current methods for assessing domain-specific knowledge and skills vary in the cognitive load they place on test-takers. As a consequence, some forms of assessment appear to engage additional domain-general cognitive resources, with consequent decreases in performance. Given the prevalence and high stakes of standardized educational assessments internationally, our findings suggest that the development and evaluation of educational assessments should extend beyond simple psychometric evaluations of validity and reliability to include consideration of the cognitive processes and abilities required for successful test performance. Specifically, the a priori specification of the knowledge, skills and abilities to be assessed must be clearly considered and explicated (e.g., the number of words that can be correctly spelled or the ability to correct erroneous spellings). Subsequent assessment design should also consider how to optimize the cognitive load placed on test-takers, to ensure that students' results are an accurate reflection of their true levels of competency. In the context of the current study, doing so would require redefinition of the abilities that tests assess (i.e., spelling vs. proofreading) or the redesign of assessments to ensure that the target knowledge or skills are accurately reflected in students' results. Although this study focused exclusively on current standardized spelling assessments, the insights generated are also able to inform principles of educational assessment design and development more broadly.

Acknowledgment—This research was funded by Internal Funding from the University of Wollongong – Australia.

REFERENCES

Alloway TP, Gregory D (2012) The predictive ability of IQ and working memory scores in literacy in an adult population. Int J Ed R 57:51–56.

Andrich D, Sheridan BS, Luo G (2005) RUMM2020: Rasch unidimensional models for measurement. Perth, WA: RUMM Laboratory.

Ashby FG, Ell SW, Valentin VV, Casale MB (2005) FROST: a distributed neurocomputational model of working memory maintenance. J Cogn Neurosci 17:1728–1743.

Baddeley A (2003) Working memory and language: an overview. J Commun Disord 36:189–208.
Best JR, Miller PH, Naglieri JA (2011) Relations between executive function and academic achievement from ages 5 to 17 in a large, representative sample. Learn Individ Differ 21:327–336.
Binder JR, Desai RH, Graves WW, Conant LL (2009) Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex 19:2767–2796.
Blair C, Razza RP (2007) Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. Child Dev 78:647–663.
Borsboom D, Mellenbergh GJ, van Heerden J (2004) The concept of validity. Psychol Rev 111:1061–1071.
Calleia AM, Howard SJ (2014) Assessing what students know: effects of assessment type on spelling performance and relations to working memory. J Student Engage Educ Matters 4:14–24.
Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci 3:201–215.
Croft AC (1982) Do spelling tests measure the ability to spell? Educ Psychol Meas 42:715–723.
Efron B (1985) Bootstrap confidence-intervals for a class of parametric problems. Biometrika 72:45–58.
Engle RW (2010) Role of working-memory capacity in cognitive control. Curr Anthropol 51:S17–S26.
Francis WN, Kucera H (1982) Frequency analysis of English usage: Lexicon and grammar. Boston: Houghton Mifflin.
Frisbie DA, Cantor NK (1995) The validity of scores from alternative methods of assessing spelling achievement. J Educ Meas 32:55–78.
Gathercole SE, Pickering SJ, Knight C, Stegmann Z (2003) Working memory skills and educational attainment: evidence from national curriculum assessments at 7 and 14 years of age. Appl Cogn Psychol 18:1–16.
Graham S, Harris KR, Chozempa BF (2002) Contribution of spelling instruction to the spelling, writing, and reading for poor spellers. J Educ Psychol 94:669–686.
Kelly AE (2011) Can cognitive neuroscience ground a science of learning? Educ Philos Theor 43:17–23.
Kirschner PA, Ayres P, Chandler P (2011) Contemporary cognitive load theory research: the good, the bad and the ugly. Comput Hum Behav 27:99–105.
Knudsen EI (2007) Fundamental components of attention. Annu Rev Neurosci 30:57–78.
Krishnan A, Williams LJ, McIntosh AR, Abdi H (2011) Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. Neuroimage 56:455–475.
McIntosh AR, Bookstein FL, Haxby JV, Grady CL (1996) Spatial pattern analysis of functional brain images using partial least squares. Neuroimage 3:143–157.
Michell J (2005) The logic of measurement: a realist overview. Measurement 38:285–294.
Murata A (2005) An attempt to evaluate mental workload using wavelet transform of EEG. Hum Factors 47:498–508.
Ogawa S, Lee TM, Kay AR, Tank DW (1990) Brain magnetic-resonance-imaging with contrast dependent on blood oxygenation. PNAS 87:9868–9872.
Paris SG (2005) Reinterpreting the development of reading skills. Read Res Quart 40:184–202.
Pearson H (2012) Issues in the assessment of spelling. Literacy Learn Middle Years 20:29–33.
Pellegrino JW (2001), Rethinking and redesigning education assessment (Research Report No. ED456136). Retrieved from Education Commission of the States website: http://www.ecs.org/clearinghouse/24/88/2488.htm.
Sampson PD, Streissguth AP, Barr HM, Bookstein FL (1989) Neurobehavioral effects of prenatal alcohol: Part II. Partial least-squares analysis. Neurotoxicol Teratol 11:477–491.
Santoro LE, Coyne MD, Simmons DC (2006) The reading-spelling connection: developing and evaluating a beginning spelling intervention for children at risk of reading disability. Learn Disabil Res Prac 21:122–133.
Strattman K, Hodson BW (2005) Variables that influence decoding and spelling in beginning readers. Child Lang Teach Ther 21:165–190.
Sweller S (1994) Cognitive load theory, learning difficulty, and instructional design. Learn Instr 4:295–312.
Szucs D, Goswami U (2007) Educational neuroscience: defining a new discipline for the study of mental representations. Mind Brain Educ 1:114–127.
van Merrienboer JJG, Sweller J (2005) Cognitive load theory and complex learning: recent developments and future directions. Educ Psych Rev 17:147–177.
Vander Wyk BC, Pelphrey KA (2011) Introduction to a special section of learning and individual differences: educational neuroscience. Learn Individ Differ 21:633–635.
Westwood P (1999) The correlation between results from different types of spelling test and children's spelling ability when writing. Aust J Learn Disabil 4:31–36.
Westwood PS (2008) What teachers need to know about teaching methods. Camberwell, VIC: ACER Press.
Whelan RR (2007) Neuroimaging of cognitive load in instructional media. Educ Res Rev 2:1–12.
Wiliam D (2003) National curriculum assessment: how to make it better. Res Pap Educ 18:129–136.
Willet L, Gardiner A (2009), Testing spelling – exploring NAPLAN. Paper presented at the Australian Literacy Educators Association Conference.
Zempleni MZ, Renken R, Hoeks JC, Hoogduin JM, Stowe LA (2007) Semantic ambiguity processing in sentence context: evidence from event-related fMRI. Neuroimage 34:1270–1279.

(Accepted 19 March 2015)
(Available online 25 March 2015)