Exp Brain Res (2013) 227:93–100
DOI 10.1007/s00221-013-3488-2

RESEARCH ARTICLE

Are reaction times obtained during fMRI scanning reliable and valid measures of behavior?

Jan Willem Koten · Robert Langner · Guilherme Wood · Klaus Willmes

Received: 11 October 2012 / Accepted: 14 March 2013 / Published online: 7 April 2013

Abstract Assuming that behavior observed during functional magnetic resonance imaging (fMRI) is comparable with behavior outside the scanner appears to be a basic tenet in cognitive neuroscience. Nevertheless, this assumption has rarely been tested directly. Here, we examined the reliability and validity of speeded performance during fMRI scanning by having the same 30 participants perform a battery of five reaction time (RT) tasks in two separate fMRI sessions and a standard laboratory (i.e., outside-scanner) session. Medium-to-high intra-class correlations between the three sessions showed that individual RT differences were conserved across sessions. Thus, for the range of tasks used, test–retest reliability and criterion validity of performance during scanning were satisfactory. Further, the pattern of between-task relations did not change within the scanner, attesting to the construct validity of performance measurements during scanning. In some tasks, however, RTs obtained from fMRI conditions were significantly shorter than those observed under normal laboratory conditions. In summary, RTs obtained during fMRI scanning appear to be largely reliable and valid measures of behavior. The observed RT speed-up during scanning might reflect task-specific interactions with a slightly different neuro-cognitive state, indicating some limits to generalizing brain–behavior relations observed with fMRI. These findings encourage further efforts in fMRI research to establish the external validity of within-scanner task performance.

Keywords fMRI · Scanner effects on performance · Test–retest reliability · Validity · Stress

Jan Willem Koten and Robert Langner contributed equally.

Electronic supplementary material The online version of this article (doi:10.1007/s00221-013-3488-2) contains supplementary material, which is available to authorized users.

J. W. Koten (✉) · K. Willmes
Neuropsychology Section, Department of Neurology,
Medical School, RWTH Aachen University,
Universitätsklinikum Aachen, Pauwelsstr. 30,
52074 Aachen, Germany
e-mail: Jan.Koten@gmx.de

J. W. Koten · G. Wood
Neuropsychology Section, Department of Psychology,
Karl Franzens University,
Graz, Austria

R. Langner (✉)
Institute of Clinical Neuroscience and Medical Psychology,
Heinrich Heine University Düsseldorf,
Düsseldorf, Germany
e-mail: r.langner@fz-juelich.de

R. Langner
Research Centre Jülich, Institute of Neuroscience and Medicine
(INM-1), Jülich, Germany

Introduction

It is often assumed that participants in neuroimaging experiments act "normally" (i.e., validly) and reliably (Koch et al. 2003). More specifically, brain–behavior relations observed during functional magnetic resonance imaging (fMRI) are usually taken to be comparable to brain–behavior relations assumed for traditional laboratory conditions or observed using different imaging modalities such as electroencephalography. Ultimately, this assumption is not testable because it is not possible to scan the brain without a scanner. Nevertheless, it is very well possible to compare the behavior of an individual in- and outside the scanner. Behavioral differences that arise from differences in environmental

conditions may indirectly reflect differential brain activity. To our knowledge, only two studies have so far investigated the effects of fMRI scanning on behavioral measures such as reaction time (RT). Both studies reported significant group mean RT differences between inside- and outside-scanner conditions (Koch et al. 2003; Assecondi et al. 2010). In contrast, group mean RT obtained from separate scanner sessions in the same individuals does not seem to differ for adults (Aron et al. 2006; Clément and Belleville 2009; Fernández et al. 2003; Fliessbach et al. 2010; Kiehl and Liddle 2003; Manoach et al. 2001; Stark et al. 2004; Wagner et al. 2005) or children (Krinzinger et al. 2011). In addition, several studies reported medium-to-high correlations between RTs recorded during separate scanning sessions (Aron et al. 2006; Fliessbach et al. 2010; Manoach et al. 2001; Stark et al. 2004).

So far it remains unclear whether and how individual differences in behavior are changed by the fMRI scanning process. The lack of research in this field is rather remarkable, because data acquisition and environmental conditions in the scanner are quite different from typical laboratory settings. For instance, most fMRI experiments are interrupted by regularly occurring pauses (resting-baseline conditions in block-design experiments or randomly occurring pauses, better known as null events, in event-related designs). Beside procedural differences between outside-scanner ("laboratory") experiments and fMRI experiments, large environmental and situational differences exist as well.

The scanner is a very loud and vibrant tool that offers only little space and limited comfort. Furthermore, for most participants, lying in a scanner is not the most everyday thing to do, and the uniqueness of the scanning procedure is emphasized by the safety measures taken to protect participants and personnel. As such it is an even more alien environment than the typical "noise-shielded, dimly lit" laboratory chamber in which behavioral data are often collected. Finally, participants are usually instructed to lie very still, with their heads being fixated using foam cushions or bite bars. All these variables influence the state of the participant and can potentially cause stress. This, in turn, may modulate brain and behavioral responses to the task at hand, which would challenge the assumption of equality between task-related brain processes in- and outside the scanner, undermining the generalizability of brain–behavior relationships studied with fMRI, referred to as external validity (Hommel et al. 2012).

The results of this experiment might be especially relevant to scientists using fMRI to uncover neural correlates of trait-like inter-individual differences (e.g., differences in personality, intelligence, expertise, clinical symptoms, or age). Of particular interest for such studies is test–retest reliability as well as criterion and construct validity of behavioral data obtained during scanning. Within this context, test–retest reliability is given when individual differences in RT are conserved across separate fMRI sessions. Criterion validity is demonstrated when RTs obtained during scanning are closely correlated with an accepted gold standard, that is, with RTs obtained outside the scanner. Construct validity is given when the relations among the dependent variables under study are preserved during scanning. Of general importance, finally, is the question of external validity, that is, the generalizability of the findings. In our context, external validity is given when mean RT is not changed by the scanning process itself.

Methods

Participants

The sample consisted of 30 male volunteers (mean age = 28.6 years, SD = 9.8) without a history of any psychiatric or neurological disorders, recruited for a previous genetic study (Koten et al. 2009) using an extended twin family design. Thus, performance during fMRI scanning has previously been analyzed in a genetic context, while performance in the outside-scanner condition has not been analyzed yet.

Our sample deviated from typical ad hoc or random samples in that it consisted of 10 family "triplets" comprising a twin pair and an additional sibling each, which may have reduced the range of inter-individual differences, as compared with a sample of completely unrelated participants. This potential range narrowing, in turn, would have biased our across-subject correlational analyses toward underestimating the true effects, thereby making these analyses only more rigorous. The study was approved by the local ethics committee (Comissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen, the Netherlands), and all participants gave their written informed consent before entering the study.

Tasks and procedure

The following five tasks were presented outside and inside the scanner (for a more detailed task description, see Koten et al. 2009):

1. Simple object categorization (CAT): Participants had to categorize visually presented objects into one of two object categories (fruits/vegetables vs. tools/kitchen utensils) via pressing one of two buttons (with their left or right index finger). A set of 48 stimuli was used for this task.
2. Arithmetic verification (ARITH). Participants had to verify visually presented one-digit addition or subtraction

problems via pressing one of two buttons (with their left or right index finger) for a correct or false result. A set of 48 stimuli was used for this task.

3. Working memory (DTC4, DTM2 and DTM4; modified from the Multidimensional Aptitude Battery (Jackson 1984; see also Vernon 1989). All three tasks were delayed-matching-to-sample tasks with a filled delay consisting of three phases: encoding, retention with distraction, and recognition. The three tasks differed in two aspects: First, the delay of the DTC4 task was occupied by the previously described CAT task, while the delays of both the DTM2 and the DTM4 task were occupied by the previously described ARITH task. Second, in the DTM2 task, two digits were to be retained, whereas in both the DTM4 and the DTC4 task, four digits were to be retained. In all tasks, participants had to verify after the delay phase if a newly presented digit was part of the previously presented two- or four-digit set. Every version of the working-memory task consisted of 32 reproductions of an experimental cycle.

Both the laboratory and fMRI experiments were designed in a way that is typical of these settings. That is, in contrast to the outside-scanner session, the experiments in the fMRI sessions were interspersed with regular pauses to allow hemodynamic activity to return to resting-baseline levels. The experimental protocol for the outside-scanner session consisted of the five tasks described above, which were presented in three blocks, each block separated by 5-min breaks. One block consisted of the three working-memory tasks presented in a nonstop fashion, while the other two blocks consisted of the CAT or ARITH task, respectively. Block and item order was randomized across participants. For the two inside-scanner sessions, experiments were identical to the ones described above, but blocks of trials were separated by regular pauses and preceded by instructions that lasted 14.4 s each (for details, see Koten et al. 2009). Further, the pseudo-randomized block and item order was identical for each participant.

All participants started with the laboratory (i.e., outside-scanner) session followed by two fMRI scanning sessions. The three sessions took place on the same day. The pause between the outside-scanner session and the first scanner session was approximately 1 hour, while the two scanner sessions were separated by 2 hours. Task and item order was identical for the two fMRI sessions.

The fixed order of measurement conditions (outside-scanner session always first) was chosen to prevent differential sequence effects on the two fMRI sessions, which might have affected the reliability and heritability estimation for task-related brain activity (cf. Koten et al. 2009). To minimize inter-session learning, we incorporated a practice session before the first experimental session. There, the same tasks with different stimuli were presented for 15 min each, during which each participant reached high proficiency (i.e., error rate <5 % in the last 20 trials of each task's practice session).

In all three sessions, the experiment was run using the software Presentation (www.neurobs.com) on standard PCs. In the outside-scanner session, stimuli were presented on a laptop PC with a 16″ flat screen; responses were given by pressing the left or right shift key on the laptop keyboard with the left or right index finger, respectively. In the inside-scanner sessions, stimuli were projected onto a large white screen attached to the head end of the scanner. Participants viewed the screen via a prism mounted on the head coil; responses were given by pressing either of two separate buttons of an MR-compatible response device using the left or right index finger.

Data analysis

Only correct responses were taken into account. In addition, responses faster than 250 ms were not considered as well as responses that took longer than three times the individual standard deviation above the individual mean RT of the respective task and condition. Extremely short RTs were considered to reflect premature (i.e., anticipatory) button presses, while extremely long RTs were considered to reflect temporary task disengagement (i.e., invalid non-task processing due to, e.g., mind wandering). Subsequently, valid RTs were averaged using the arithmetic mean and considered for further statistical analysis using SPSS.

Test–retest reliability and criterion validity were assessed by calculating intra-class correlations (ICCs) using the conservative absolute agreement criterion, in which the time and participant components were treated as random factors (McGraw and Wong 1996). ICCs test the hypothesis that the absolute performance level as well as its covariance structure do not differ between experimental conditions. We consider this an important assumption to be fulfilled when the external validity of a within-scanner performance measure is evaluated. In principle, large differences in absolute agreement between outside- and inside-scanner performance might seriously undermine the assumption that the same cognitive functions are measured, questioning the external validity of fMRI experiments.

Construct validity was assessed with a repeated-measures analysis of variance (ANOVA), testing for an interaction between task type and session (outside vs. inside the scanner). Finally, external validity was explicitly assessed (apart from using the aforementioned strict ICC approach) by the same ANOVA, testing for the main effect of session. Additionally, we used paired t tests to examine whether significant differences in mean RT existed between measurements in- and outside the scanner for all the three possible

combinations (outside scanner with scanner run A; outside scanner with scanner run B; scanner run A with scanner run B) for all five experiments. Since the three working-memory experiments included a distraction task to fill the delay, they yielded two performance measures each (i.e., for the primary and secondary task), leading to a total $3 \times 8 = 24$ t tests. As we expected to find no differences between outside- and inside-scanner sessions, p-values were not corrected for multiple comparisons. This made the analyses more conservative, since it raised the chance to detect undesirable differences, that is, to falsify our prediction. Finally, supplementary analyses examined the time course of RT in each experiment for the outside-scanner condition to test for any remaining learning effects that might have occurred despite the preceding practice sessions (see Supplementary Material for details).

Results and discussion

Group-level statistics for mean RT are given in Table 1 (for other parameters of intra-individual RT distributions, see Table S1 in the Supplementary Material). The grand average analysis across all tasks revealed high ICCs (coefficient range = 0.79–0.91) between the three sessions, suggesting that individual RT differences were conserved irrespective of the measurement condition under study. The latter was also the case when individual RT experiments were assessed for their reliability and criterion validity (ICC range = 0.69–0.91) except for the CAT task, which showed poorer reliability in all conditions (ICC range = 0.42–0.64). This reduced reliability is probably due to the small inter-individual variability, as reflected by the low SD values (see Table 1). Overall, these data suggest that RT obtained during scanning shows sufficient test–retest reliability (correlation between the two fMRI sessions) and criterion validity (correlation between the outside-scanner session and the first fMRI session).

Group mean RTs are shown in Fig. 1. The ANOVA yielded a significant effect of session [$F(1, 29) = 10.61$, $p = 0.003$] and task [$F(7, 203) = 110.03$, $p < 0.001$] but no significant interaction between them [$F(7, 203) = 1.16$)]. The absence of this interaction indicates that relations among the tasks were not changed by the scanning process, attesting to the construct validity of RT measurements inside the scanner. The main effect of session revealed that speed increased in the fMRI session (cf. grand means in Table 1). A supplementary ANOVA including RT raw data (without outlier removal) yielded the same result (for details, see supplementary results and discussion as well as Table S1). Subsequent paired samples t tests showed that the global speed increase from outside- to first inside-scanner session was driven by 5 out of 8 individual tasks (cf. Table 1 and Fig. 1). In contrast, when comparing the first with the second fMRI session, no significant global difference emerged ($F = 1.57$). Such a stability of mean RT over scanner sessions was also observed in previous studies with adults (Aron et al. 2006; Clément and Belleville 2009; Fernández et al. 2003; Fliessbach et al. 2010; Kiehl and Liddle 2003; Manoach et al. 2001; Stark et al. 2004; Wagner et al. 2005) and children (Krinzinger et al. 2011).

To examine a potential change of the speed–accuracy trade-off from outside- to inside-scanner sessions, we analyzed different measures of performance accuracy. Trials with anticipatory, extremely slow, or missing responses were generally very rare and therefore not further analyzed, with group mean percentages ranging across tasks and sessions from 0 to 0.3 % for anticipatory responses; from 0.6 to 2.1 % for extremely delayed responses; and from 0 to 1.9 % for missing responses (see supplementary Table S2). Errors (i.e., wrong button presses) occurred somewhat more frequently, with group mean error rates ranging across tasks from 1.5 to 7.8 % during the outside-scanner session; from 3.1 to 6.3 % during the first inside-scanner session; and from 2.2 to 5.7 % during the second inside-scanner session (see Table S2). A repeated-measures ANOVA of arcsine-transformed error rate yielded no significant main effect of session [$F(1, 29) = 2.19$] but a significant effect of task [$F(7, 203) = 10.44$, $p < 0.001$] and a significant session × task interaction [$F(7, 203) = 3.93$, $p = 0.001$]. Subsequent paired samples t tests showed that this interaction was driven by a significant decrease in error rate from outside- to first inside-scanner session for one task [CAT: $t(29) = 4.46$, $p < 0.001$] and a significant increase for two tasks [DTC4_2: $t(29) = -2.21$, $p = 0.035$; DTM2_2: $t(29) = -2.15$, $p = 0.040$]. Taken together, accuracy was not generally affected by the scanning situation but showed some minor yet inconsistent changes, which rules out a global explanation of (task-specific) RT improvements during scanning based on shifts of the speed–accuracy criterion.

In summary, RT was rather stable across the two separate scanner sessions. There was, however, a global though not completely consistent increase in speed from the outside-scanner condition to the first inside-scanner condition. Together with previous studies that reported a significant decrease in speed (Koch et al. 2003; Assecondi et al. 2010), our data demonstrate that mean RT can significantly change in the fMRI scanner. Although the direction of change appears to depend on the specific task under study, the available evidence undermines the external validity of RT measures collected during fMRI scanning, given that results obtained under "normal" laboratory conditions constitute the accepted gold standard. However, since the direction of change does not seem to be systematic across different tasks and studies (Koch et al. 2003; Assecondi et al. 2010; Hommel et al. 2012; Chouinard et al. 2008), more research

Table 1 Within- and between-session performance measures for each task (1 outside- and 2 inside-scanner sessions)

Task/session	Within-session measures					Between-session measures			
	Reaction time (ms)				ΔRT split half	ICC		ΔRT(*p*-value)	
	M	*SD*	*Md*	Skew	*p*-value	Outside	Inside 1	Outside	Inside 1
CAT									
Outside	801	142	805	0.10	0.76				
Inside 1	774	110	775	0.52		0.63		0.179	
Inside 2	712	115	684	0.37		0.43	0.42	<0.001*	0.007*
ARITH									
Outside	1,091	242	1,006	0.30	0.38				
Inside 1	1,027	227	1,004	0.34		0.85		0.006*	
Inside 2	997	245	944	0.44		0.79	0.87	<0.001*	0.171
DTC4-2									
Outside	789	177	735	0.79	0.27				
Inside 1	734	157	718	1.04[§]		0.77		0.008*	
Inside 2	696	135	669	0.40		0.62	0.69	<0.001*	0.073
DTC4-3									
Outside	817	225	698	0.75	0.85				
Inside 1	784	218	680	0.62		0.90		0.073	
Inside 2	772	212	720	0.65		0.83	0.91	0.046*	0.470
DTM2-2									
Outside	1,145	239	1,118	0.28	0.55				
Inside 1	1,088	255	999	0.44		0.81		0.041*	
Inside 2	1,068	242	1,050	0.11		0.75	0.87	0.013*	0.395
DTM2-3									
Outside	731	181	667	0.67	0.69				
Inside 1	683	179	662	0.64		0.76		0.035*	
Inside 2	684	195	647	0.61		0.61	0.79	0.123	0.968
DTM4-2									
Outside	1,139	251	1,091	0.29	0.23				
Inside 1	1,065	235	1,000	0.47		0.81		0.005*	
Inside 2	1,049	227	1,042	0.18		0.70	0.84	0.007*	0.517
DTM4-3									
Outside	810	218	716	0.94[a]	0.17				
Inside 1	785	204	712	0.69		0.87		0.209	
Inside 2	786	217	785	0.26		0.74	0.87	0.403	0.978
Grand Mean									
Outside	915	190	553	0.64	0.29				
Inside 1	867	181	524	0.66		0.88		0.003*	
Inside 2	845	182	514	0.25		0.74	0.86	0.004*	0.219

This table reports reaction time (RT) mean, standard deviation, median, and skewness per task and session. Next, *p*-values from paired *t* tests comparing mean RT in the first and last half of each task ("ΔRT split half") in the outside-scanner condition are given. Finally, intra-class correlation (ICC) coefficients and *p*-values from paired *t* tests comparing mean RT between experimental conditions are given. The grand mean represents average values across all 8 tasks

Task abbreviations modified after the Multidimensional Aptitude Battery (Jackson 1984): *CAT* object categorization, *ARITH* arithmetic verification, *DTC4-2/3* 2nd (secondary task)/3rd (primary task) phase of the 4-digit memory task using object categorization as secondary task, *DTM2-2/3* 2nd (secondary task)/3rd (primary task) phase of the 2-digit memory task using arithmetic verification as secondary task, and *DTM4-2/3* 2nd (secondary task)/3rd (primary task) phase of the 4-digit memory task using arithmetic verification as secondary task

* Significant at $p < 0.05$

[a] Skewness deviates significantly from 0 (i.e., skewness/$SE_{skewness} > 2$)

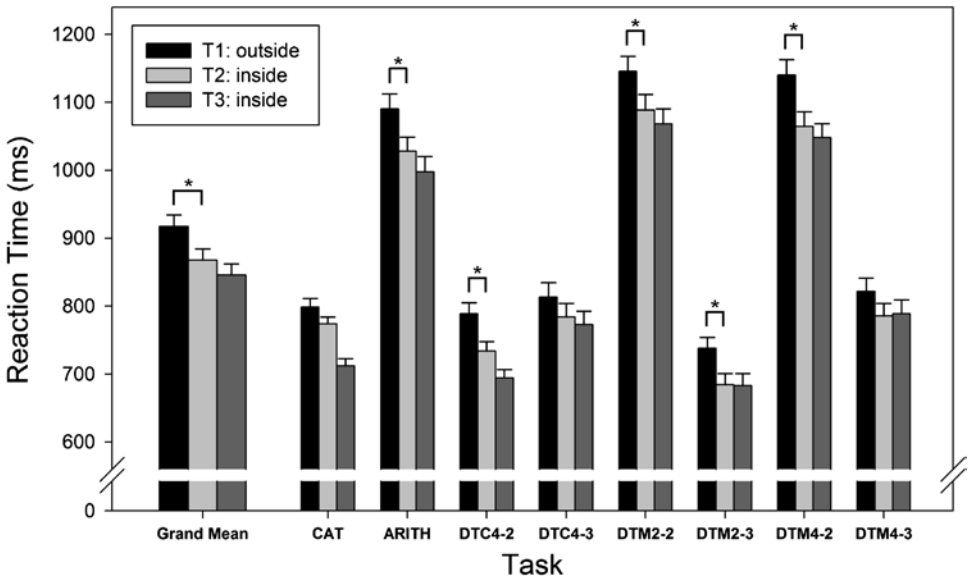

Fig. 1 Group mean reaction time (RT) for all tasks and sessions. Grand mean represents mean RT averaged across all eight tasks. Significant ($p < 0.05$) RT reductions from outside- to inside-scanner sessions are marked by an *asterisk* (*). T1 = outside-scanner session; T2/T3 = first/second inside-scanner session. For a list of task abbreviations, please see Table 1

is needed to generalize this conclusion. This is all the more important, because earlier findings were based on rather small or, as in the present case, male-only samples.

Although future studies are required to systematically examine and disentangle the factors driving these scanning-related performance changes, we will briefly discuss some potential mechanisms. A mechanism that may come to mind first, despite the countermeasures taken, is learning/practice: Since we were unable to randomize or counterbalance the order of measurement conditions (i.e., the outside-scanner session always came first), it might be assumed that the between-session performance change was a consequence of increased proficiency acquired over sessions. In search of evidence for this assumption, we looked for speed changes over the course of each outside-scanner condition, based on the notion that, according to the power law of practice (Newell and Rosenbloom 1981), learning-induced changes should become most evident early during practice. The mean time courses, however, were remarkably stable across the outside-scanner session (see supplementary results and Figure S1). This absence of within-session improvements (which also held true for both inside-scanner sessions; cf. Figure S2) argues against learning as the major mechanism underlying between-session differences, but of course, it does not completely rule out the possibility that between-session consolidation contributed to the observed improvements. Finally, the between-task variability in performance change across sessions also makes an explanation based on inside–outside hardware differences unlikely, given the high similarity of the tasks' perceptual and motor demands.

In our view, the performance differences between measurement environments point to a role of the scanner-specific situation (see Hommel et al. 2012, for a similar view). As mentioned above, there are several differences between the standard laboratory and the scanner environment that may affect the functional state of the participant. Factors such as noise, vibration, narrowness, and movement restrictions can act as stressors (Hockey 1984), and a slight stress-induced increase in arousal during the scanner sessions might have beneficial effects on performance by enhancing the availability of attentional resources and/or decreasing the demand for effortful compensatory arousal regulation (Fischer et al. 2008). Also, performance impairments arising from a depletion of attentional resources, mind wandering, or both over time (Langner et al. 2010; Langner and Eickhoff 2012) might have been mitigated by the regular breaks (i.e., resting-baseline blocks) implemented in the fMRI-specific task schedules. Finally, cognitive task performance in the fMRI scanner could have benefitted from the lying position of the participants, because posture control, on which less demand is put during lying than sitting, can interfere with cognitive processing (Chong et al. 2010; Harley et al. 2006).

As a result, it appears desirable that, whenever possible, performance should be measured while brain activity is being acquired in the scanner, rather than correlating task-related brain activity with performance in the same task measured outside the scanner. Otherwise, true brain–behavior relationships could be underestimated in tasks in which performance is affected by the scanning procedure. Conversely, relationships between brain activity and behavioral or subjective measures taken outside the scanner could also be overestimated when these measures share variance with any effect of the scanner-specific situation on brain activity. The concurrent measurement of performance and brain activity does not, however, solve the problem of reduced external validity arising from potential fMRI scanning effects on performance and—by extrapolation—on brain activity. This also implies that comparisons between

brain–behavior relationships observed using different imaging modalities non-concurrently (i.e., in separate sessions) should only be made with caution. Therefore, the external validity of within-scanner performance measurements should be tested on a more regular basis in fMRI research. Ultimately, beyond the issue of external validity with respect to the gold standard "laboratory task," there is an increasing need to examine—and potentially improve—the *ecological* validity of fMRI experiments to gain further insights into how the brain works in real life (Snow et al. 2011).

Conclusion

Overall, our results, derived from five different cognitive tasks yielding eight measures of speeded performance, indicate that speeded behavior observed during fMRI scanning can be sufficiently reliable for the meaningful statistical analysis of individual differences. In addition, inter-individual differences in response speed as well as inter-task relations remained remarkably well conserved when inside- and outside-scanner conditions were compared. Scanning-related increases in response speed, however, suggest that behavior in the scanner is not always quite the same as compared with standard laboratory conditions but may reflect enhanced task engagement. We conclude that RTs obtained during fMRI scanning can be used for reliable and valid inferences on brain–behavior relations. However, external validity may be weakened by the additive impact of situation-specific factors of the fMRI environment, which might induce a somewhat different neuro-cognitive state. Therefore, combining on-line performance measurement during scanning with an additional outside-scanner measurement (i.e., the accepted gold standard) appears to be the optimal approach for examining generalizable brain–behavior relationships.

References

Aron AR, Gluck MA, Poldrack RA (2006) Long-term test-retest reliability of functional MRI in a classification learning task. NeuroImage 29(3):1000–1006. doi:10.1016/j.neuroimage.2005.08.010

Assecondi S, Vanderperren K, Novitskiy N, Ramautar JR, Fias W, Staelens S, Stiers P, Sunaert S, Van Huffel S, Lemahieu I (2010) Effect of the static magnetic field of the MR-scanner on ERPs: evaluation of visual, cognitive and motor potentials. Clin Neurophysiol 121(5):672–685. doi:10.1016/j.clinph.2009.12.032

Chong RK, Mills B, Dailey L, Lane E, Smith S, Lee KH (2010) Specific interference between a cognitive task and sensory organization for stance balance control in healthy young adults: visuospatial effects. Neuropsychologia 48(9):2709–2718. doi:10.1016/j.neuropsychologia.2010.05.018

Chouinard PA, Morrissey BF, Kohler S, Goodale MA (2008) Repetition suppression in occipital-temporal visual areas is modulated by physical rather than semantic features of objects. NeuroImage 41(1):130–144. doi:10.1016/j.neuroimage.2008.02.011

Clément F, Belleville S (2009) Test-retest reliability of fMRI verbal episodic memory paradigms in healthy older adults and in persons with mild cognitive impairment. Hum Brain Mapp 30(12):4033–4047. doi:10.1002/hbm.20827

Fernández G, Specht K, Weis S, Tendolkar I, Reuber M, Fell J, Klaver P, Ruhlmann J, Reul J, Elger CE (2003) Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. Neurology 60(6):969–975

Fischer T, Langner R, Birbaumer N, Brocke B (2008) Arousal and attention: self-chosen stimulation optimizes cortical excitability and minimizes compensatory effort. J Cogn Neurosci 20(8):1443–1453. doi:10.1162/jocn.2008.20101

Fliessbach K, Rohe T, Linder NS, Trautner P, Elger CE, Weber B (2010) Retest reliability of reward-related BOLD signals. NeuroImage 50(3):1168–1176. doi:10.1016/j.neuroimage.2010.01.036

Harley C, Boyd JE, Cockburn J, Collin C, Haggard P, Wann JP, Wade DT (2006) Disruption of sitting balance after stroke: influence of spoken output. J Neurol Neurosurg Psychiatry 77(5):674–676. doi:10.1136/jnnp.2005.074138

Hockey GRJ (1984) Varieties of attentional state: The effects of the environment. In: Parasuraman R, Davies DR (eds) Varieties of attention. Academic Press, Orlando, pp 449–483

Hommel B, Fischer R, Colzato LS, van den Wildenberg WP, Cellini C (2012) The effect of fMRI (noise) on cognitive control. J Exp Psychol Hum Percept Perform 38(2):290–301. doi:10.1037/a0026353

Jackson DN (1984) Multidimensional Aptitude Battery. Research Psychologists Press, Port Huron

Kiehl KA, Liddle PF (2003) Reproducibility of the hemodynamic response to auditory oddball stimuli: a six-week test-retest study. Hum Brain Mapp 18(1):42–52. doi:10.1002/hbm.10074

Koch I, Ruge H, Brass M, Rubin O, Meiran N, Prinz W (2003) Equivalence of cognitive processes in brain imaging and behavioral studies: evidence from task switching. NeuroImage 20(1):572–577

Koten JW Jr, Wood G, Hagoort P, Goebel R, Propping P, Willmes K, Boomsma DI (2009) Genetic contribution to variation in cognitive function: an fMRI study in twins. Science 323(5922):1737–1740. doi:10.1126/science.1167371

Krinzinger H, Koten JW, Hennemann J, Schueppen A, Sahr K, Arndt D, Konrad K, Willmes K (2011) Sensitivity, reproducibility, and reliability of self-paced versus fixed stimulus presentation in an fMRI study on exact, non-symbolic arithmetic in typically developing children aged between 6 and 12 years. Developmental neuropsychology 36(6):721–740. doi:10.1080/87565641.2010.549882

Langner R, Eickhoff SB (2012) Sustaining attention to simple tasks: a meta-analytic review of the neural mechanisms of vigilant attention. Psychol Bull. doi:10.1037/a0030694

Langner R, Willmes K, Chatterjee A, Eickhoff SB, Sturm W (2010) Energetic effects of stimulus intensity on prolonged simple reaction-time performance. Psychol Res 74(5):499–512. doi:10.1007/s00426-010-0275-6

Manoach DS, Halpern EF, Kramer TS, Chang Y, Goff DC, Rauch SL, Kennedy DN, Gollub RL (2001) Test-retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. Am J Psychiatry 158(6):955–958

McGraw KO, Wong SP (1996) Forming inferences about some intraclass correlation coefficients. Psychol Meth 1:30–46

Newell A, Rosenbloom PS (1981) Mechanisms of skill acquisition and the law of practice. In: Anderson JR (ed) Cognitive skills and their acquisition. Erlbaum, Hillsdale, pp 1–55

Snow JC, Pettypiece CE, McAdam TD, McLean AD, Stroman PW, Goodale MA, Culham JC (2011) Bringing the real world into the fMRI scanner: repetition effects for pictures versus real objects. Scientific reports 1:130. doi:10.1038/srep00130

Stark R, Schienle A, Walter B, Kirsch P, Blecker C, Ott U, Schäfer A, Sammer G, Zimmermann M, Vaitl D (2004) Hemodynamic effects of negative emotional pictures - a test-retest analysis. Neuropsychobiology 50(1):108–118. doi:10.1159/000077948

Vernon PA (1989) The heritability of measures of speed of information processing. Personality Individ Differ 10:573–576

Wagner K, Frings L, Quiske A, Unterrainer J, Schwarzwald R, Spreer J, Halsband U, Schulze-Bonhage A (2005) The reliability of fMRI activations in the medial temporal lobes in a verbal episodic memory task. NeuroImage 28(1):122–131. doi:10.1016/j.neuroimage.2005.06.005