NeuroImage 73 (2013) 260–264

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Review

Changes in brain structure during learning: Fact or artifact? Reply to Thomas and Baker

R. Douglas Fields

Nervous System Development and Plasticity Section, The Eunice Kennedy Shriver National Institute of Child Health and Human Development, 35 Lincoln Drive, Bldg. 35, Room 2A211, Bethesda, MD 20892, USA

ARTICLE INFO

Article history:
Accepted 31 August 2012
Available online 7 September 2012

Keywords:
White matter plasticity
Learning and memory
Environmental enrichment
MRI
DTI
Activity-dependent myelination
Neuroimaging
Plasticity

ABSTRACT

In their review in this issue, Thomas and Baker question the validity of longitudinal human neuroimaging studies that have claimed to demonstrate structural plasticity. This commentary identifies problems with some of the arguments raised in their review and suggests that there is strong evidence, from both animal and human studies, that experience can alter brain structure.

Contents

Introduction

In 1962, Rosenzweig et al. (1962), looking for biochemical changes in the brain during learning, serendipitously observed that the physical structure of the cerebral cortex was altered by experience. Rats that experienced an enriched environment, containing mazes, toys, and affording increased social interaction, had a larger cerebral cortex than control rats not exposed to the stimulating environment (Fig. 1). This finding was made by using the simplest of instruments, a scalpel and a precision laboratory balance. It was made using a straight-forward experimental design, a comparison of differences between an experimental and control population, in which the subjects were assigned to groups randomly and the analysis was done without knowledge of the experimental condition. This pioneering finding launched decades of research to identify the cellular changes in brain tissue responsible for the increased brain mass associated with learning and environmental experience. Changes in many cellular components of the brain have been found after learning, including neuronal numbers, synaptic density, dendritic complexity, axon sprouting, glial numbers, cell morphology, cell differentiation, myelination, and changes in vascular cells, reviewed elsewhere (Fields, 2008, 2011; Fu and Zuo, 2011; Holtmaat and Svoboda, 2009; Zatorre et al., 2012).

Fifty years later, Blumenfeld-Katzir et al. (2011) reported structural changes in the brain of rats after training in a Morris water maze that

E-mail address: fieldsd@mail.nih.gov.

1053-8119/$ – see front matter © 2012 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.neuroimage.2012.08.085

MEAN BRAIN WEIGHTS OF EXPERIMENTAL AND CONTROL GROUPS (IN MILLIGRAMS)

Strain	N (Pair)	Sensory Cortex		Total Dorsal Cortex		Total Cortex		Subcortex I		Subcortex II		Total Brain	
		ECT	IC	ECT	IC	ECT	IC	ECT	IC	ECT	IC	ECT	IC
S_1	11	90	82	335	320	663	626	1222	1202	894	896	1557	1522
S_8	9	93	93	399	395	738	731	1374	1385	1035	1048	1773	1779
K	12	104	100	431	405	707	688	1310	1338	1035	1055	1741	1743
RCH	10	106	103	457	443	809	774	1390	1385	1037	1053	1847	1828
All	42	98	94	405	390	726	701	1320	1324	999	1011	1725	1713

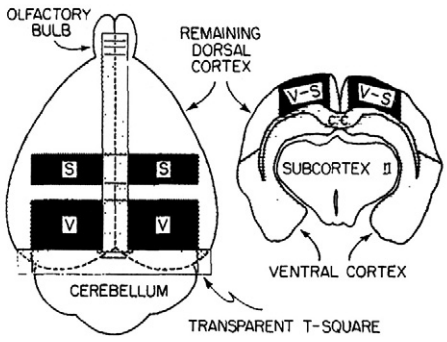

Fig. 1. Experience and training change the physical structure of the brain. Data from Rosenzweig et al. (1962) showing that in all four strains of rats that were studied, environmental complexity and training (ECT) increased the weight of the cerebral cortex compared with isolated controls (IC) reared in standard cages. The increase in cortical mass was largest in the dorsal and total cortex (arrows); 4% larger on average than in controls. The inset shows how the brain was dissected for analysis. V = visual cortex, S = somatosensory cortex.

were observed by MRI and confirmed by histology. Microscopic analysis showed statistically significant differences in neurons, glia, and myelin proteins consistent with changes detected by MRI in the appropriate brain regions.

The review by Thomas and Baker (2013-this issue) express misgivings about the simple two-sample experimental design and questions the large and diverse body of MRI data using various other experimental designs that find structural changes in the human brain after training. The thrust of the argument is that the studies showing structural changes in the human brain after training are unsound because MRI methodology lacks sufficient precision to detect the small changes produced relative to the experimental noise and errors introduced by procedures of analysis, and in their view, the authors of studies reporting positive findings have not designed their experiments correctly or analyzed their data with appropriate statistical methods.

> "*We conclude that the current literature on training-dependent plasticity in adult humans does not provide unequivocal evidence for training-dependent structural changes and more rigorous experimentation and statistical testing is required* (p. 9)."

The authors then offer advice about how such studies should be designed and analyzed.

Proof by consensus

A deep fault underlying the approach taken in this review is the "appeal to the majority" flawed logic that is applied.

> "*…of the 14 studies that tested for correlations between behavior and structural changes, only half reported evidence for significant effects* (p. 7)."

> "*With a few exceptions… the vast majority of the studies we reviewed simply present statistical parametric maps of the differences between pre- and post-training scans* (p. 5)."

> "*In only five* [of 20] *studies…could we find any indication that the appropriate statistical tests to identify training-related structural changes were conducted* (p. 4)."

A faulty study does not invalidate the finding of a well-constructed study. How does this tabulation of studies failing to meet the authors' specific criteria undermine the conclusions of the studies that do satisfy the authors' conditions? Glossing over or mixing together such important differences among studies as the study objectives, experimental design, type of learning task involved, duration of training, age of subjects, time-points of analysis, measurement methods, replication, etc., makes cross-comparisons ambiguous.

The authors adopt the terminology of statistics, "robustness, strength, power, rigor, correlation, signal-to-noise, proof," but they use the words in the vernacular sense, rendering the arguments rhetorical and ambiguous. Statistics do not "prove" a hypothesis. Logic dictates that the only conclusion possible from a statistical analysis arises when the data disprove a hypothesis; typically that there is no difference between two or more sampled populations. The statistical analysis defines the probability that the differences in measurements obtained by sampling could have occurred by chance. A study that fails to reject the null hypothesis, e.g., Thomas et al. (2009), does not lead to any conclusion, other than a difference could not be detected. This outcome does not invalidate the conclusions of other studies that succeed.

Absence of evidence

The facts presented are that of 20 studies in the literature reporting positive effects, one study failed to find an influence of learning/training on human brain structure examined with MRI (Thomas et al., 2009). Thomas and Baker's position is that this negative study calls into question all the others reporting positive findings. Accepting this conclusion requires making two assumptions: (1) that the negative finding in the Thomas et al. (2009) study is correct, and (2) that the experimental conditions in the Thomas et al. (2009) study are equivalent to all the other studies showing positive effects, thereby justifying invalidating the conclusions of all the others.

Neither assumption is certain. If the latter assumption is not correct, the criticism presented would be a hasty generalization from a special case. Close examination of the Thomas et al. (2009) study reveals several reasons to question generalizing from this study. The age of the subjects is older than in most of the other studies (mean

age of 32.5 years), nearly a decade older than subjects in many of the other studies showing positive results. Skill learning and brain plasticity diminishes with age (Boyke et al., 2008). The period of training was shorter than many studies, comprising only six 25 min sessions spread over 14 days. The training task used was less challenging than many of the tasks used in other studies reporting positive findings. The training involved learning to operate a joystick when the direction of the controls is reversed. However, individuals encounter similar situations in normal experience: backing up a car, operating a boat with an out-board motor, moving the stage on a microscope, interpreting movements seen in a reflection. It seems possible that improving on this task might not require structural remodeling of brain tissue. If so, the results of Thomas et al. (2009) would be correct, but this would not undermine positive findings in other studies using more demanding types of learning, including musical performance, juggling, and various difficult physical skills and mental tasks where positive findings are reported.

Alternatively, the Thomas et al. (2009) study could have failed to detect a difference between experimentals and controls when in fact a difference exists; that is, the results are a false negative (committing a type II error). There are some aspects of the experimental design that could have weakened the ability to detect a difference had there been one. The sample size is small (only 12 individuals). The control group is heterogeneous and data from two different experimental designs were combined without necessary statistical analysis to justify doing so. Contrary to what is stated in Thomas and Baker's review, the study by Thomas et al. (2009) had variable controls and reference images, ranging from base-line images taken immediately before training to three months prior to training. Four of the 12 subjects had control images taken a mean of 44.5 days before training. One experimental design utilized three time points for sampling (base-line, non-treated control, and post training; all separated by two week intervals). The other experimental design used four time points for sampling (base-line, non-treated, a second non-treated sample taken up to three months later, and a post training scan). The data from both designs were combined under the untested assumption that there was no difference between the second and third scans in the four-sample design. The conclusion of the Thomas et al. (2009) study is that the results they obtain differ depending on which images are used as references because of uncontrolled variance. If the baseline is not stable or the experimental design is heterogeneous, this potential outcome (failing to detect a difference) would be more likely; especially if the effects of the treatment are small or variable. Thus, accepting Thomas and Baker's argument built upon this single study requires making two critical assumptions that are uncertain and abandoning the reasoning upon which hypothesis testing and the scientific method are based.

Favoring type II errors

The criticism in their review is a warning about the perils of committing a type I error (concluding a difference exists when there is none), and it offers detailed suggestions about how to design experiments to reduce type I errors in MRI studies of learning, but this well-intentioned and useful advice about reducing noise, dealing with small signals, how to design control groups, how much replication is necessary, the sensitivity and precision of the instruments, random errors and sources of systematic errors, artifacts, signal-to-noise issues, and covariance with other factors, is not pertinent. A false positive (type I error) is just as wrong as a false negative (type II error). The consequences of either error depend on context, and the threshold of committing a type I or type II error is controlled by setting the criterion probability value accepted for significance. Using a $p<0.001$ will increase the likelihood of committing a type II error (false negative); using a $p<0.05$ will increase the likelihood of committing a type I error (a false positive). This will obtain regardless of the experimental design or measurement method. Indeed, these conclusions are mathematically defined clear interpretations regardless of what experiments were performed, equipment used, or hypothesis that motivated the data sampling that generated the numbers.

Bias and uncontrolled variables and measurement error

That experimental data collection can be undermined by bias or subject to systematic errors is well appreciated. Certainly alternative interpretations for differences found to be statistically significant are always possible and important to consider, but this alternative interpretation is resolved by additional experiments to test a new specific hypothesis—either with new data or by reanalyzing the data sorted according to the new factor under consideration—not by the approach taken in their review. The criticism that differences between controls and experimental conditions may reflect more head motion in the controls because these subjects were less motivated than the experimental subjects, for example, is not a basis for rejecting the statistically defined difference between groups. Rather it requires accepting that the difference found by the experiment cannot be due to chance, but instead posits an alternative interpretation for the difference, which is testable. Unless the alternative factor can be identified in the experiment and eliminated, or the factor is examined separately in a different experiment, attacking studies on this basis is condemnation by innuendo.

> "*…the changes in FA reported in participants who took part in integrative body-mind training rather than relaxation therapy…could in principle arise from physiological changes* (*e.g. cardiac or respiration*) *induced in one of the groups following the training, rather than the specific meditation technique* (p. 7)."

This alternative hypothesis could be tested experimentally, but without doing so this argument begs the question in rejecting the results of the original findings. Similarly:

> "*…at each stage of the data processing stream, from preprocessing to statistical analyses, significant biases can be introduced inadvertently, and these can give rise to spurious changes in brain structure* (p. 9)."

If experimental errors and artifacts are distributed randomly, the finding of a statistically significant difference between groups justifies rejecting the null hypothesis. If these errors are not distributed randomly or biases influence the data collection or analysis, then the fundamental assumptions of any statistical analysis (or any experiment) are violated and no conclusion is possible. However, unless these assumptions are shown to have been violated, the results cannot be rejected by criticism that such things are possible.

Size of the effect

The authors argue that the changes expected in brain structure during learning are too small to see in human brain imaging.

> "*The effect size reported in some human studies is very small relative to the size of the voxels* (p.2)."

> "*It is not clear that human MRI, with typically 1 mm^2 voxels, can detect the type of microscopic structural changes reported in animal studies* (p. 2)."

However, the magnitude of an effect is a different question from whether the data obtained from sampling groups are significantly different, and this determination requires a different method of statistical analysis. Paradoxically, the authors accept that the MRI data showing changes in brain structure in animal studies are correct (e.g., Blumenfeld-Katzir et al. (2011), but the authors do not explain how insufficient resolution is not a concern in the rat brain, which

is roughly the size of the human eyeball, but undermines data obtained from imaging the much larger human brain.

> "*Given that such effect sizes are many times smaller than the sampling frequency of the method, these results need to be carefully evaluated and interpreted with caution* (p.2)."

Accepting the authors' position that summary data less than the unit of measure is suspect would invalidate all types of averaging, including signal averaging or averaging data in populations. A difference of 1.9 and 2.7 in mean number of children/family between two populations can be real and the difference can have significant social and financial consequences, regardless of the fact that a fractional child does not exist: "effect size smaller than the sampling." Moreover, this criticism, and many others in the review, could be applied equally to MRI data in many different contexts. MRI is used in thousands of publications concerning human disease, development, as well as medical diagnosis. It is not clear why the criticisms are applied selectively to the body of research on human brain imaging associated with learning.

Controls

> "*Despite the importance of collecting control data, four of the 19 studies we identified did not include any control condition… making it impossible to assess the training-dependence of any reported effects* (p. 4)."

This statement is not correct. These four studies used a repeated measures design in which changes were monitored in each individual after training, so that each subject was its own control, much as blood pressure changes might be measured in a person before and after taking medication for hypertension. Many of these studies said to lack a control also used correlation analysis showing strong and highly significant changes in MRI data with training and performance.

Correlation between behavioral changes and brain structure

The authors would require that accepting the conclusion that structural changes in the brain accompany learning, requires that behavioral measures of learning are also included in the study and found to correlate with the MRI changes. This thinking is a broken syllogism that fails to recognize the difference between necessity and sufficiency. There are many reasons to expect that behavioral measures (improved performance) might not require structural changes or that such a correlation with MRI might not be found. Many factors are involved in performance of a complex task. Some are necessary but not sufficient for performance, and the same is likely to be the case for structural changes in the brain. Secondly, it is not clear which behavioral measure should correlate with a structural change observed—performance peak, speed of learning, amount of improvement, accuracy, practice time or specific aspects of the skill. Third, there are technical and biological reasons why MRI changes are seen in different brain regions at different times and ages, and associated with different aspects of tasks. Such correlations, or lack of them, are interesting and important questions in the field of learning, but this is a different issue from whether or not the MRI data are a measurement artifact.

The authors argue that the structural changes detected by MRI must be "specific to a given task not a general effect of any training (p. 4)." However, different parts of the brain are engaged at different times and for different purposes, and likely develop differently in people depending on their prior skill and learning strategy. Spatial information is moved from the hippocampus to the neocortex during slow wave sleep, for example. The cerebellum is involved in learning and in many other brain functions beyond motor control. As individuals develop specific skills the task becomes more automated and the cognitive process engages cortical regions to different extents or in different regions than when the task was a new challenge. These are questions at the forefront of information processing and learning, which MRI studies of human learning together with fMRI are beginning to answer. To presume the answers to these questions and then use them as a premise for testing the conclusion that structural changes in the brain accompany learning begs the question.

Conclusion

The review advances an opinion on an important subject that has been widely reviewed (Draganski and May, 2008; Fields, 2011; May, 2011; Zatorre et al., 2012), but the approach taken cannot lead to a certain conclusion, and the thesis advanced is undermined by internal inconsistency. The authors argue both ways: that structural changes in the brain accompany learning in animals and possibly in humans, but they simultaneously reject the evidence for the conclusion.

> "*Such hemisphere-specific training is often used in animal model studies, yielding evidence that is highly compelling* (p. 6)."

> "*We do not mean to suggest that the other studies reporting structural changes are invalid or provide no evidence for training-dependent structural plasticity, just that the strength of the evidence is limited and alternative interpretations of the apparent effects are possible* (p. 7)."

> "*We do not suggest that the training protocols used in the human studies do not induce structural changes in the adult brain. Nor do we imply that MRI cannot be used to detect training-dependent structural changes in the adult brain. On the contrary, the two animal studies…described earlier provided compelling evidence for the feasibility of MRI-based techniques to detect training-related structural changes in adult animals* (p. 34)."

That the experimental evidence for cellular changes in human brain during learning is less compelling at present than in animal studies is to be expected for the obvious reasons of limitations on sampling human brain tissue. Long-term potentiation (LTP), the cellular model of learning at a synaptic level, was first described in rabbits by Bliss and Lomo (1973), but evidence of LTP in the human brain did not begin to emerge until 33 years later (Cooke and Bliss, 2006). Evidence for experience-dependent structural changes in the brain of experimental animals, including myelination, is not diminishing; on the contrary strong evidence continues to accumulate (Makinodan et al., in press). The rationale for rejecting the findings from animal studies with respect to the cellular mechanisms of learning in the human brain is missing from the analysis.

Skepticism is vital in science. Publication bias toward positive results is appropriate and necessary because negative findings provide no logical conclusion. If through experimental bias incorrect results get published, as in the memory transfer experiments of the 1960s, they will be repudiated eventually (e.g., Luttges et al., 1966); this is the normal and healthy way that science should proceed. The future will no doubt find some studies in the field of brain imaging and learning flawed or wrong, but it is experimental data and adherence to the logic of the scientific method of hypothesis testing that brings answers. This holds whether using an MRI machine or a balance to investigate how experience alters brain structure.

Acknowledgment

This work was supported by funds for intramural research at NICHD.

References

Bliss, T.V., Lomo, T., 1973. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. J. Physiol. 232, 331–356.

Blumenfeld-Katzir, T., Pasternak, O., Dagan, M., Assaf, Y., 2011. Diffusion MRI of structural brain plasticity induced by a learning and memory task. PLoS One 6, e20678.

Boyke, J., Driemeyer, J., Gaser, C., Buchel, C., May, A., 2008. Training-induced brain structure changes in the elderly. J. Neurosci. 28, 7031–7035.

Cooke, S.F., Bliss, T.V.P., 2006. Plasticity in the human central nervous system. Brain 129, 1659–1673.

Draganski, B., May, A., 2008. Training-induced structural changes in the adult human brain. Behav. Brain Res. 192, 137–142.

Fields, R.D., 2008. White matter in learning, cognition and psychiatric disorders. Trends Neurosci. 31, 361–370.

Fields, R.D., 2011. Imaging learning: the search for a memory trace. Neuroscientist 17, 185–196.

Fu, M., Zuo, Y., 2011. Experience-dependent structural plasticity in the cortex. Trends Neurosci. 34, 177–187.

Holtmaat, A., Svoboda, K., 2009. Experience-dependent structural synaptic plasticity in the mammalian brain. Nat. Rev. Neurosci. 10, 647–658.

Luttges, M., Johnson, T., Buck, C., Holland, J., McGaugh, J., 1966. An examination of "transfer of learning" by nucleic acid. Science 151, 834–837.

Makinodan, M., Rosen, K.M., Ito, S., and Corfas, G., in press. A critical period for social experience-dependent oligodendrocyte maturation and myelination. Science.

May, A., 2011. Experience-dependent structural plasticity in the adult human brain. Trends Cogn. Sci. 15, 475–482.

Rosenzweig, M.R., Krech, D., Bennett, E.L., Diamond, M.C., 1962. Effects of environmental complexity and training on brain chemistry and anatomy: a replication and extension. J. Comp. Physiol. Psychol. 55, 429–437.

Thomas, C., Baker, C.I., 2013. Teaching an adult brain new tricks: A critical review of evidence for training-dependent structural plasticity in humans. NeuroImage 73, 225–236 (this issue).

Thomas, A.G., Marrett, S., Saad, Z.S., Ruff, D.A., Martin, A., Bandettini, P.A., 2009. Functional but not structural changes associated with learning: an exploration of longitudinal voxel-based morphometry (VBM). Neuroimage 48, 117–125.

Zatorre, R.J., Fields, R.D., Johansen-Berg, H., 2012. Plasticity in gray and white: neuroimaging changes in brain structure during learning. Nat. Neurosci. 15, 528–536.