Psychiatry Research: Neuroimaging 212 (2013) 167–174

Contents lists available at SciVerse ScienceDirect

Psychiatry Research: Neuroimaging

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

An application of item response theory to fMRI data: Prospects and pitfalls

Michael L. Thomas [a], Gregory G. Brown [a,b,*], Wesley K. Thompson [a], James Voyvodic [c], Douglas N. Greve [d], Jessica A. Turner [e], Daniel H. Mathalon [f], Judith Ford [f], Cynthia G. Wible [g], Steven G. Potkin [h], FBIRN

[a] University of California, Department of Psychiatry, La Jolla, San Diego, CA 92093, United States
[b] VA San Diego Healthcare System, VISN-22, Mental Illness, Research, Education, and Clinical Center, San Diego, CA 92161, United States
[c] Duke University Medical Center, Brain Imaging and Analysis Center, Durham, NC 27710, United States
[d] Massachusetts General Hospital, Department of Radiology, Cambridge, MA 02114, United States
[e] The Mind Research Network, Albuquerque, NM 87106, United States
[f] University of California, Department of Psychiatry and San Francisco VA Medical Center, San Francisco, CA 94143, United States
[g] Department of Psychiatry, Harvard Medical School and Brockton VAMC, Cambridge, MA 02301, United States
[h] University of California, Department of Psychiatry, Irvine CA, 92697, United States

ARTICLE INFO

Article history:
Received 2 July 2012
Received in revised form
11 January 2013
Accepted 29 January 2013

Keywords:
Functional MRI
Item response theory
Bayesian analysis
Task design

ABSTRACT

When using functional brain imaging to study neuropsychiatric patients an important challenge is determining whether the imaging task assesses individual differences with equal precision in healthy control and impaired patient groups. Classical test theory (CTT) requires separate reliability studies of patients and controls to determine equivalent measurement precision with additional studies to determine measurement precision for different levels of disease severity. Unlike CTT, item response theory (IRT) provides estimates of measurement error for different levels of ability, without the need for separate studies, and can determine if different tests are equivalently difficult when investigating differential deficits between groups. To determine the potential value of IRT in functional brain imaging, IRT was applied to behavioral data obtained during a multi-center functional MRI (fMRI) study of working memory (WM). Average item difficulty was approximately one standard deviation below the ability scale mean, supporting the task's sensitivity to individual differences within the ability range of patients with WM impairment, but not within the range of most controls. The correlation of IRT estimated ability with fMRI activation during the task recognition period supported the linkage of the latent IRT scale to brain activation data. IRT can meaningfully contribute to the design of fMRI tasks.
Published by Elsevier Ireland Ltd.

1. Introduction

Over the past several decades, item response theory (IRT; Lord and Novick, 1968; Rasch, 1960) has become the preferred methodology for the study of test and item characteristics. Yet, IRT has only rarely been applied in neuropsychological research, and almost never in published functional brain imaging studies. In this paper, we discuss some of the practical issues researchers are likely to confront when applying these techniques to functional brain imaging studies. This demonstration is accomplished by applying IRT to behavioral data obtained during a multi-center functional MRI (fMRI) study of working memory. Readers wishing a more general discussion of IRT should consult introductory texts (e.g., de Ayala, 2009; Embretson and Reise, 2000), review papers (e.g., Reise and Waller, 2009; Thomas, 2011), and technical resources (e.g., Baker and Kim, 2004; van der Linden and Hambleton, 1997).

1.1. Motivation for using IRT in functional brain imaging

Although interesting fMRI studies are being performed under behaviorally unconstrained conditions (Meda et al., 2012), most of the studies in the fMRI literature have used cognitive challenge tasks to probe patterns of brain-activation. Behavioral contributions to the design of fMRI tasks have focused almost exclusively on the validity of the task as an apparent assessment of cognitive neuroscience domains of interest. Once the content validity of items is determined, item properties such as difficulty and discriminating power are assumed, often implicitly, to be equivalent across items. When item difficulty is considered, it typically enters through the manipulation of independent variables, such as memory load or stimulus visibility, that alter the marginal probability of a correct response over subgroups of items (Huang et al., 2006; Potkin et al., 2009). However, item difficulty needs to be considered when designing

* Corresponding author at: VA San Diego Healthcare System, Psychology Service (116B), 3350 La Jolla Village Dr., San Diego, CA 92161, United States.
E-mail address: gbrown@ucsd.edu (G.G. Brown).

0925-4927/$ - see front matter Published by Elsevier Ireland Ltd.
http://dx.doi.org/10.1016/j.pscychresns.2013.01.009

brain activation probes in order to avoid ceiling and floor effects, especially when studying groups of subjects who perform at different ability levels (Gur et al., 1992). Difficulty should be matched across cognitive challenge probes in order to support the attribution of differential brain response to the different neurocognitive systems that the probes were designed to evoke (Gur et al., 1992; Snyder et al., 2011; Spitzer et al., 1996).

These initial applications of psychometric ideas to the design of brain activation tasks were not developed within an explicit psychometric framework, although the principles of classical test theory (CTT) often seem to be assumed. Today, IRT offers an accessible, advanced set of tools for establishing the precision and accuracy of individual items (see Embretson and Hershberger, 1999). IRT models involve both individual person parameters and individual item parameters scaled along the same latent dimension. This focus results in an explicit model of item and person characteristics that are differentiated while remaining linked to each other through a parametric equation. Separation of person and item parameters allows for invariance of item characteristics across groups and individuals that differ in ability (Lord, 1980), and provides an explicit rationale for the use of different items to assess the same neurocognitive system in diverse groups of patients (e.g., adaptive testing methods). IRT also permits the assessment of item information (similar to the concept of reliability) and standard error at specific points along the ability spectrum, whereas CTT would require different reliability studies along arbitrarily quantitized intervals of ability. Measurement precision can be determined independently for groups and individuals with different ability levels, as often occurs in functional brain imaging studies of clinical groups (e.g., Brown and Eyler, 2006).

The primary purpose of using IRT in imaging research is to evaluate item properties in order to ensure that tests are measuring intended neurocognitive constructs with appropriate difficulty to detect individual differences in latent ability; a precise approach to the ideas advocated by Gur et al. (1992). Unfortunately, there are several obstacles to using IRT in imaging studies; most notably, the typically large subject samples required to estimate IRT parameters and questions whether or not the latent abilities estimated in IRT are related to brain activation. A test of IRT's practical utility in imaging research is needed.

1.2. An application of IRT to an fMRI study of working memory

Data come from the East Coast Traveling Subjects (ECTS) study performed by the Function Biomedical Informatics Research Network (FBIRN). The aim of the study was to assess the multi-site reliability of functional imaging data before embarking on a larger multi-center study of schizophrenia patients. Participants were administered a working memory task (WMT) designed to detect differential patterns of brain activation of healthy volunteers and schizophrenia patients with working memory impairment. The WMT is a forced-choice delayed visual recognition memory test, permitting the separate detection of brain processes involved in stimulus encoding, memory maintenance, and target recognition. The task was presented in the magnet while images sensitive to blood oxygen level dependence (BOLD) signals were acquired (see Buxton, 2002).

To model WMT item characteristics, we consider nested versions of a general IRT model where N examinees respond to J items. Let $X_{ij}=x_{ij}$ denote the observed response for the i^{th} examinee to the j^{th} item, where $x_{ij}=1$ if the response is correct and 0 otherwise.[1] The probability of a correct response is approximated by a logistic function of subject ability (θ_i), item difficulty (β_j), item discrimination (α_j), and item guessing (γ_j) parameters

$$P(X_{ij}=x_{ij}|\beta_j,\alpha_j,\gamma_j,\theta_i)=\gamma_j+(1-\gamma_j)\frac{e^{\alpha_j(\theta_i-\beta_j)}}{1+e^{\alpha_j(\theta_i-\beta_j)}} \quad (1)$$

Eq. (1) is commonly referred to as a three-parameter logistic (3-PL; Birnbaum, 1968) model. The θ_i parameter reflects the subject's standing on the underlying ability that is required for accurate item responding (e.g., memory). It is an unobservable characteristic of the examinee that may also be referred to as a latent factor or trait. The β_j or item difficulty parameter makes it more or less probable that an examinee of a given ability level will provide a correct response. The α_j or discrimination parameter reflects the weight or relevance of the underlying ability dimension to the probability of a correct response. The γ_j or lower-asymptote parameter conveys the probability that an examinee with infinitely low ability will correctly respond (often guessing).

IRT models range from simple to complex in both scope and ease of application. For imaging researchers hoping to use IRT in their work, it is first necessary to consider what combination of freely estimated item parameters can be viably attained from available data. The answer is due, in part, to characteristics of items, but also practical issues related to sample size. It is challenging to collect large samples in imaging research due to cost, time, and access barriers associated with scanning equipment. In the current study, for instance, item responses and imaging data were collected for 18 participants over nearly 6 months of multisite collaboration at a cost of approximately $1000 per scanning session, per site. This reduced number of examinees – which is common in cognitive and imaging research – can annul the beneficial large sample properties of maximum likelihood estimators (see Baker and Kim, 2004). It is well known, for example, that samples sizes should range from several hundred to several thousand participants for simple to complex IRT models respectively (de Ayala, 2009; Reckase, 2009). Sample sizes of ≤ 50 can result in biased parameter estimates or fail to converge, even for simple models (Lord, 1968). Unstable or biased estimates of item characteristics associated with small sample sizes are especially troublesome for maximum likelihood and least squares estimators. Later, we discuss the use of Bayesian estimators with constraining prior information to improve model convergence and fit.

As with most imaging data sets, the WMT data structure is a transpose of the typical psychometric data set. That is, whereas psychometric data are characterized by a greater number of subjects than items, the current data are characterized by a greater number of items than subjects. This is seen as a problem in IRT, because whereas subjects are typically modeled with just a single parameter, items are modeled with multiple parameters. As the ratio of subjects to items grows smaller, it becomes increasingly difficult to accurately estimate item parameters.

Fortunately, there may be characteristics of items that, when combined with certain types of estimation procedures, can overcome this challenge. It is generally known that traction in parameter estimation can be gained by constraining item characteristics to single, group values (see Wainer and Wright, 1980). This strategy works well when individual item parameters show only minor deviations from the group average, and do not significantly deteriorate model fit when held constant. A more general, less stringent framework for this strategy comes from hierarchical Bayesian modeling, where individual items are assumed to be drawn from common distributions (Levy, 2009). If the properties of these distributions (e.g., shape, mean, and variance) are known, or can be assumed based on experimental control and prior theory, limitations in the estimation of item properties from observed data can be mitigated. The WMT, like most cognitive tasks used in imaging research, makes use of highly

[1] A more complex model that included a site difficulty parameter was also investigated. The model poorly converged and did not fit the data better than models excluding site effects. Consequently no site term was included in the model.