MULTIPLE SCLEROSIS JOURNAL | MSJ

Topical Review

The use and quality of reporting of propensity score methods in multiple sclerosis literature: A review

Mohammad Ehsanul Karim, Fabio Pellegrini, Robert W Platt, Gabrielle Simoneau, Julie Rouette and Carl de Moor

Multiple Sclerosis Journal

2022, Vol. 28(9) 1317–1323

DOI: 10.1177/1352458520972557

Abstract

Background: Propensity score (PS) analyses are increasingly used in multiple sclerosis (MS) research, largely owing to the greater availability of large observational cohorts and registry databases.

Objective: To evaluate the use and quality of reporting of PS methods in the recent MS literature.

Methods: We searched the PubMed database for articles published between January 2013 and July 2019. We restricted the search to comparative effectiveness studies of two disease-modifying therapies.

Results: Thirty-nine studies were included in the review, with most studies (62%) published within the past 3 years. All studies reported the list of covariates used for the PS model, but only 21% of studies mentioned how those covariates were selected. Most studies used PS matching (72%), followed by PS adjustment (18%), weighting (15%), and stratification (3%), with some overlap. Most studies using matching or weighting reported checking post-PS covariate imbalance (91%), although about 45% of these studies relied on *p* values from various statistical tests. Only 25% of studies using matching reported calculating robust standard errors for the PS analyses.

Conclusions: The quality of reporting of PS methods in the MS literature is sub-optimal in general, and in some cases, inappropriate methods are used.

Keywords: Comparative effectiveness, causal inference, disease-modifying therapies, multiple sclerosis, observational study, propensity score

Date received: 30 June 2020; revised: 27 September 2020; accepted: 28 September 2020.

Correspondence to:
ME Karim
School of Population & Public Health, University of British Columbia, 2206 East Mall, Vancouver, BC V6T 1Z3, Canada.
ehsan.karim@ubc.ca

Mohammad Ehsanul Karim
School of Population & Public Health, University of British Columbia, Vancouver, BC, Canada/ Centre for Health Evaluation and Outcome Sciences, University of British Columbia, Vancouver, BC, Canada

Fabio Pellegrini
Biogen International GmbH, Zug, Switzerland

Robert W Platt
Julie Rouette
Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada/Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, Montreal, QC, Canada

Gabrielle Simoneau
Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada/Biogen Canada, Mississauga, ON, Canada

Carl de Moor
Biogen, Cambridge, MA, USA

Introduction

Multiple sclerosis (MS) is a chronic demyelinating disease of the central nervous system with no known cure. A number of disease-modifying therapies (DMTs) are available in the market to help reduce the severity and frequency of MS relapses and development of new lesions in the brain with the goal of slowing down the progression of the disease.[1] Important variations exist, however, in the clinical efficacy, tolerability, and safety profile of these DMTs.[2]

Randomized controlled trials (RCTs) are the gold standard for establishing causality in comparative effectiveness studies. However, they are not always feasible. The availability of various registries and large multicentre MS cohorts thus opened the opportunity to answer research questions that are impractical to investigate using RCTs.[3–6]

The estimation of causal treatment effects in observational studies is challenging, mainly because treatment is not assigned at random. Hence, these sources can be prone to biases, such as selection bias and confounding by indication. Statistical techniques for observational data analyses currently offer several tools to reduce bias due to confounding. Notably, the PS approaches, which allow to adjust for the influence of confounders, have been increasingly used in the MS literature.[5]

There is a current gap in the literature reviewing the use of PS methods in MS. This article aims to address this gap by critically reviewing the MS literature using PS methods in terms of methodological use and reporting.

Methods

We searched the PubMed database for articles published in English between January 2013 and July

2019. Articles were included if the study (1) pertained to MS patients, (2) was a comparative effectiveness study of DMTs, (3) evaluated the effectiveness between two arms (or treatment groups), and (4) used PS methods. Extracted data included study background, objectives, PS methodology, and statistical approach. Additional methods, including the necessary technical background on PS methods to fully appreciate the observations in this review, can be found in the Supplementary Methods.

Observations

A total of 64 articles were identified during the title and abstract screening of which 39 were retained for data extraction (see Supplementary Results for additional results). Table 1 summarizes the characteristics of the 39 comparative effectiveness studies included in the review. Between 2013 and 2019, we observed a gradual increase in studies using PS methods in the MS literature. All studies were cohort studies, using either MS-specific databases or registries (82%), study-specific databases (3%), claims databases (12%), or general hospital-based databases (3%). Most studies were conducted across multiple sites (82%), while the remaining studies were single-site studies. The total sample size ranged from 92 to 12,042 patients, with a median of 951 patients.

The most common DMT was fingolimod in 56% of studies, followed by natalizumab in 38% of studies. A total of 90% of studies specified at least one outcome as primary. Approximately 54% of these studies reported at least one continuous primary outcome, 37% reported at least one binary primary outcome, and 46% reported at least one time-to-event primary outcome. The following primary outcomes were most commonly reported: annualized relapse rate (39%), time to first relapse (18%), time to disease progression (21%), proportion of patients who experienced a relapse (13%), treatment persistence (11%), and time to treatment discontinuation (8%). About 85% of the studies reported at least one secondary outcome.

PS estimation

Table 2 shows the general characteristics of the reporting of the PS model and analysis. Most studies (90%) assessed covariate imbalances before conducting the PS analysis. Of those, 86% noted pre-PS imbalances. About 13% of studies did not report how the PS model was estimated. Most studies (87%) used logistic regression to estimate the PS, with 5% of studies resorting to statistical model selection. On average, seven covariates were used to construct the PS (range: 3–16). All studies reported a list of these variables, but only 21% of studies reported how this list was determined: of those, 75% used expert opinion and 25% used statistical tests. The following variables were most commonly used: age, sex, disease duration, number of relapses in the 12 months prior to baseline, Expanded Disability Status Scale (EDSS) score at baseline, and treatments at baseline. Magnetic resonance imaging-related measures (presence of gadolinium-enhancing or cerebral T2 lesions) were used in 28% of the studies. About 5% of the studies included post-baseline variables (annualized relapse rate) in the PS model. Most studies (72%) used PS matching, 15% used weighting, and 18% used PS regression adjustment. Only one study used stratification (with quintiles).

Matching

Table 3 presents the characteristics of the 28 studies that used PS matching as the method of analysis. The majority of studies (54%) used 1:1 matching. One study did not report the matching ratio. Two studies did not report the algorithm used to match patients. Most studies (93%) used greedy nearest neighbor matching. Among these studies, caliper widths varied widely, with 7% of studies using a caliper along with exact matching on a given covariate (baseline EDSS and disease duration). About 29% of studies used a caliper but did not report the chosen caliper. Most studies implemented matching without replacement (64%), although 32% of studies did not indicate whether matches were selected with or without replacement. On average, the reduction between the initial and matched sample size was 46% (range: 2%–89%). Most studies (75%) used standardized mean differences (SMDs) to check post-PS imbalances, and 43% of studies reported *p* values based on various tests. About 4% of studies did not state if or how post-PS imbalances were checked. Post-PS imbalances were noted in 32% of studies, but only one study took action to rectify the imbalances. Only 25% of the studies reported using robust standard errors in the outcome analysis.

Weighting

Six of the 39 studies (15%) used weighting. Half of these studies estimated the average treatment effect among the treated, 17% estimated the average treatment effect, and 33% did not report the causal effect of interest. About 66% of studies checked post-PS imbalances, among which 75% of studies used SMDs, and 25% used *p* values (*t*-test, Kolmogorov–Smirnov, or chi-square). One study used trimming at 2.5% of

Table 1. Characteristics of the 39 included studies.

Variables	No studies/total (%)
Publication year	
2013	1/39 (3)
2014	1/39 (3)
2015	8/39 (21)
2016	5/39 (13)
2017	7/39 (18)
2018	14/39 (36)
2019 (up to July)	3/39 (8)
Data source[a]	
MS-specific registry or database	32/39 (82)
Study-specific database	1/39 (3)
Claims database	5/39 (13)
Healthcare administrative database	1/39 (3)
No. of patients enrolled	
Median [Q1–Q3]	951 [563–2557]
Mean [Min.–Max.]	2,012 [92–12,042]
No. of treated patients	
Median [Q1–Q3]	428 [191–756]
Mean [Min.–Max.]	897 [37–11,657]
No. of control patients	
Median [Q1–Q3]	455 [268–985]
Mean [Min.–Max.]	831 [49–6605]
Minimum treatment exposure or follow-up period eligibility criteria	
Yes	26/39 (67)
No	13/39 (33)
Treatments under consideration[b]	
Interferon-β	10/39 (26)
Glatiramer acetate	3/39 (8)
Dimethyl fumarate	7/39 (18)
Fingolimod	22/39 (56)
Teriflunomide	2/39 (5)
Natalizumab	15/39 (38)
Alemtuzumab	2/39 (5)
Mixed treatments[c]	13/39 (33)
Other[d]	9/39 (23)
No. primary outcomes	
Single	23/39 (59)
Multiple	12/39 (31)
Not specified[e]	4/39 (10)
No. secondary outcomes	
None	6/39 (15)
1	7/39 (18)
2–5	18/39 (46)
>5	8/39 (21)

Max.: maximum, Min.: minimum, No.: number, Q1: first quartile, Q3: third quartile.
[a]MS-based registries or databases were iMED (4), MSBase (12), Tysabri Observational Program (2), CLIMB (1), and country-specific registries (France, USA, Austria, Canada, Denmark, Italy, Germany, Sweden, Switzerland).
[b]Studies that compared two formulations of the same DMT (e.g. interferon beta-1a and -1b) counted once toward that DMT.
[c]Studies where one comparison group included multiple DMTs, for example, BRACE (Betaseron (interferon beta-1b), Rebif (interferon beta-1a), Avonex (interferon beta-1a), Copaxone (Glatiramer acetate), and Extavia (interferon beta-1b)).
[d]No treatment (5), cladribine (1), and rituximab (1).
[e]Four studies had secondary outcomes but did not specify the primary outcome.

Table 2. Reporting of propensity score analysis in 39 studies.

Characteristics	No studies/total (%)
Baseline imbalances noted pre-PS	
Yes	30/39 (77)
No	5/39 (13)
Not reported	4/39 (10)
PS model	
Logistic regression	34/39 (87)
Not reported	5/39 (13)
No. of variables to estimate the PS	
Median [Q1–Q3]	7 [6–9]
Mean [Min.–Max.]	7 [3–16]
Inclusion of non-baseline variables in the PS	
Yes	2/39 (5)
No	37/39 (95)
PS method[a]	
Matching	28/39 (72)
Weighting	6/39 (15)
Stratification	1/39 (3)
Adjustment	7/39 (18)

Max.: maximum, Min.: minimum, No.: number, PS: propensity score.
[a]Three studies used two methods.

each tail as a sensitivity analysis; no other study reported using trimming or truncation. Only one study used stabilized weights. Only 33% of the studies reported using robust standard errors.

Discussion

PS methods are particularly appealing for comparative effectiveness studies in MS given the increasing availability of non-experimental data. The objective of this study was to review the use and quality of reporting of PS methods in the recent MS literature. We identified 39 comparative effectiveness studies using PS methods published between 2013 and 2019. We summarized the extracted data in terms of the general characteristics of the studies, the estimation of the PS, and the resulting PS adjusted analysis. We observed some good practices that are followed by MS researchers using PS approaches; for example, the list of confounders used to construct the PS model was always reported, most of the studies that used PS matching reported the algorithm used to match patients, and post-PS imbalance checking was conducted by most studies, often with SMDs. However, many other essential aspects of the reporting of PS methodology were not optimal within the MS literature. Here, we summarize the found gaps in the use and reporting of PS methods and derive a few general and MS-specific recommendations.

Table 3. Characteristics of 28 studies that used matching within the 39 studies reviewed.

Characteristics	No. studies/28 (%)
Matching ratio[a]	
1:1	15/28 (54)
1:2	3/28 (11)
1:3	3/28 (11)
Variable[b]	7/28 (25)
Not reported	1/28 (4)
Matching algorithm	
Greedy[c]	26/28 (93)
Not reported	2/28 (7)
Caliper	
0.1 or less	15/28 (54)
0.2	1/28 (4)
Proportion of SDs[d]	4/28 (14)
Not reported	8/28 (29)
Use of replacement[e]	
With replacement	2/28 (7)
Without replacement	18/28 (64)
Not reported	9/28 (32)
Reduction of matched sample size	
< 25%	5/28 (18)
25–50%	12/28 (43)
50–75%	7/28 (25)
> 75%	4/28 (14)
Method to assess balance[f]	
SMD	21/28 (75)
p value[g]	12/28 (43)
Other[h]	1/28 (4)
Not reported	3/28 (11)
Post-PS imbalances noted	
No	18/28 (64)
Yes	9/28 (32)
Not reported	1/28 (4)

PS: propensity score, SMD: standardized mean difference.
[a]One study used 1:1 and 1:2 matching in two separate analyses.
[b]Variable ratios ranged from 1:1 to 1:10.
[c]One study combined greedy matching with other algorithms (Kernel, radius).
[d]0.01 SD (SD), 0.25 SDs (1), 0.3 SDs (1), 0.5 SDs (1).
[e]One study used both with and without replacement to construct two separate matched cohorts.
[f]Some studies used more than one method to assess balance.
[g]*p* values were based on *t*-tests, McNemar tests, Wilcoxon signed-rank tests, chi-square tests, Fisher exact tests, Mann–Whitney *U* tests or logistic regression. One study did not report which tests were used to derive *p* values.
[h]Comparison of summative and average distances of the PS in the two treatment groups between the unmatched and matched samples.

Areas of poor reporting

In many studies, the reporting of how the PS model was derived and estimated often lacked details. Although all studies reported the list of confounders that were included in the PS model, 79% of studies did not state how this list was obtained. Authors should provide enough information about why they included the selected confounders, for example, from the opinion of an expert in the domain or following findings in other studies. Over-reliance on statistical tests for variable selection should be discouraged in general because different sample sizes will result in different covariates being selected based on *p* values.[7,8] Most studies reported fitting the PS model with a logistic regression but did not indicate how the model was specified, that is, with only main-effects, or whether polynomial terms and interactions were included in the model. Also, basic verifications of the fitted PS model were rarely reported; visual assessments of the overlap of the distribution of the fitted PS by treatment groups should be performed, for example, with boxplots, histograms, or density plots.[9] Assessments of the positivity assumption should be explicitly reported, for instance, by inspecting the fitted PS directly or by noting confounders with no occurrence or low frequency in one of the treatment groups. Unfortunately, zero-cell issues were clearly noticeable in 13% of the studies under consideration.

Our review found a high proportion of PS matching studies that checked post-PS imbalances (96% of studies). This assessment is an essential aspect of quality reporting in PS methods. However, we found that 25% of studies that used PS matching did not report SMDs. Generally, the strategy to correct for those post-PS imbalances is to first improve the PS model. If balance on some covariates still cannot be achieved, then the imbalanced covariates should be included in the outcome regression model.[10] However, it was not clear if this was done for the studies with a high imbalance (e.g. studies with SMD $>$ 0.2). Also, post-PS imbalances were less often checked in studies that used weighting (66%); SMDs should be used to assess balance in the weighted samples as well.

Most studies did not report how they handled missing data in their analysis. Among the few studies which reported missing data analyses, the majority presented ad hoc or single imputation approaches, which suffer from known statistical limitations[11] (see Supplementary Discussion). Suboptimal handling of missing data may affect the results from a PS analysis as it would affect any other types of analyses.

Post-baseline covariates

Adjustment with post-baseline covariates in PS models is highly discouraged, as confounders can only affect the treatment assignment if they are measured before or at the time of treatment decision. Post-baseline variables are likely an effect of the treatment or worse, a mediator—adjusting for either of such variables has serious statistical consequences.[12] One possible explanation could be that the researchers considered the post-baseline variable as a proxy for an important known but unmeasured pre-baseline covariate and hence adjusted for it to get an approximate treatment effect estimate.[13] We found that 5% of studies included such post-baseline covariates in their PS models to adjust for potential imbalances between the treated and control groups. However, even in this case, the direction of bias for the treatment effect estimate would be hard to guess without additional information regarding the relationship between that important unmeasured variable and the proxy variable.[7] Results from such analyses should be interpreted with caution. MS DMTs usually require sustained treatment strategies (i.e. exposure over time), whereas standard PS approaches are not capable of handling time-varying variables, and more useful for point treatment strategies.[14] Unfortunately, only a few published works within MS literature have considered applying more appropriate longitudinal time-dependent analyses approaches for sustained treatment strategies, which have the ability to appropriately adjust for post-baseline variables as well as time-varying exposure.[15,16]

Comparison with other disease areas

We compared our findings to similar reviews in other disease areas to highlight features specific to MS (see Supplementary Methods). Our comparison highlighted a few encouraging practices in MS research. The use of *p* values to check post-PS covariate balance is lower in MS than in other disease areas, but it remains high. It is generally recommended to check covariate balance with SMDs instead of *p* values from conventional statistical tests. Our comparison also highlighted some problems specific to MS. For example, our review had the lowest median number of covariates included in the PS model. One argument could be that MS researchers were presumably more careful in selecting confounders for their analysis. One could also argue that there are fewer factors in MS that confound the relationship of interest compared to other disease areas. However, it is not possible to assure the readers whether either was the case, as most reviewed studies did not report how they have

selected the covariates to include in the PS model. However, researchers argue that some notable confounders or outcome measures helpful in guiding MS treatment decisions are not adequately or routinely measured in MS-based registries and cohorts.[17] Even though the omission of important confounders likely leads to biased treatment effect estimates (e.g. due to apparent violation of the conditional exchangeability assumption), most studies in our review (64%) did not formally assess the influence of unmeasured confounding, for example, via Rosenbaum bounds.

Reproducibility and generalizability

Lack of clarity in the reporting of PS analyses may have serious implications for research in MS. Failure to clearly report the methodology utilized may increase the risk that the results of such studies are misinterpreted. For example, we found that 77% of all studies did not indicate if standard errors were calculated with a robust approach. Consequently, in those studies, the reliability of the reported confidence intervals for the treatment effect is unclear. Lack of transparency and relevant details in a manuscript also makes it difficult to reproduce the results. Inadequate reporting may, in turn, affect subsequent research and, ultimately, clinical practice.

In our review, we found that 39% of the studies deleted at least 50% of the study subjects while matching. Such drastic reductions in sample sizes may impact the precision of the results. Furthermore, selectively excluding a subset of the target population may also impact the generalizability or external validity of the study. In such analyses, the resulting underlying target population may have been reduced to a very specific sub-population, and the estimated treatment effect may not be generalizable to the original target population. From an opposing perspective, sacrificing external validity (deleting a large number of subjects) could be seen as a necessary step for achieving internal validity. In such a situation, for the sake of generalizability, it might be advisable not to use matching as the primary analysis but rather to explore alternative PS approaches that do not require such extreme measures.[18,19] Researchers should not feel obliged to use one specific approach and should perform necessary sensitivity analyses to validate the study findings. A related issue for such observational studies is whether this target population is clearly reported. In the absence of a clear definition of the target population, the concept of generalizability is lost, regardless of the quality of the PS methodology.

Limitations

Our review has some limitations. First, it was challenging to evaluate precisely how the PS analyses were performed based on the published material. If the researchers did not clearly report the necessary PS model development and diagnostic steps in their article, it was not possible to assess whether they correctly executed the analysis. For example, details about graphical inspections of the overlap of the distributions of the estimated PS between the two treatment groups were rarely shown.[20] Second, the PS estimation methodology was often reported with some important details omitted (e.g. especially when reported as a sensitivity analysis), preventing the extraction of relevant information that would be helpful for understanding the analytical choices. For example, it was often difficult to identify if the PS model was formed with main-effect terms only or if polynomial or interaction terms were also considered.

Conclusion

PS methods are increasingly used in comparative effectiveness studies in the MS literature. While our review highlights some good practices in the use and reporting of PS methods in MS, there are rooms for improvement in designing methodologically rigorous studies and reporting crucial information in order to enhance reproducibility and generalizability. The development of MS-specific guideline for the use and reporting of PS methods would be helpful to aid in the appropriate applications of PS methods and ensure transparency.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: M.E.K. is supported by the Michael Smith Foundation for Health Research Scholar award and holds research grants from the Natural Sciences and Engineering Research Council of Canada and BC SUPPORT Unit. He has received consultancy fees from Biogen over the past 3 years. F.P. is a Biogen employee and owns stock of the company. R.W.P. has received consultancy fees from Biogen. G.S. has received consultancy fees from Biogen during the preparation of this manuscript before becoming a Biogen employee before its completion. G.S. owns Biogen stocks. J.R. has received consultancy fees from Biogen. C.d.M. is a Biogen employee and owns stock of the company.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This review was funded by Biogen.

ORCID iD

Gabrielle Simoneau https://orcid.org/0000-0001-9310-6274

Supplemental Material

Supplemental material for this article is available online.

References

1. Cohen JA and Rudick RA. *Multiple sclerosis therapeutics*: Cambridge: Cambridge University Press, 2011.
2. Gajofatto A and Benedetti MD. Treatment strategies for multiple sclerosis: When to start, when to change, when to stop? *World J Clin Cases* 2015; 3: 545–555.
3. Kalincik T and Butzkueven H. Observational data: Understanding the real MS world. *Mult Scler* 2016; 22(13): 1642–1648.
4. Zaratin P, Comi G and Leppert D. 'Progressive MS–macro views': The need for novel clinical trial paradigms to enable drug development for progressive MS. *Mult Scler* 2017; 23(12): 1649–1655.
5. Trojano M, Tintore M, Montalban X, et al. Treatment decisions in multiple sclerosis—Insights from real-world observational studies. *Nat Rev Neurol* 2017; 13(2): 105–118.
6. Cohen JA, Trojano M, Mowry EM, et al. Leveraging real-world data to investigate multiple sclerosis disease behavior, prognosis, and treatment. *Mult Scler* 2020; 26(1): 23–37.
7. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol* 2019; 34: 211–219.
8. Schneeweiss S and Suissa S. Advanced approaches to controlling confounding in pharmacoepidemiologic studies. *Pharmacoepidemiology* 2019: 1078–1107.
9. Shrier I, Pang M and Platt RW. Graphic report of the results from propensity score method analyses. *J Clin Epidemiol* 2017; 88: 154–159.
10. Nguyen T-L, Collins GS, Spence J, et al. Double-adjustment in propensity score matching analysis: Choosing a threshold for considering residual imbalance. *BMC Med Res Methodol* 2017; 17: 78.
11. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* 2009; 338: b2393.
12. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70: 41–55.
13. Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J R Stat Soc Ser A: G* 1984; 147: 656–666.
14. Young JG, Vatsa R, Murray EJ, et al. Interval-cohort designs and bias in the estimation of per-protocol effects: A simulation study. *Trials* 2019; 20: 552.
15. Spelman T, Freilich J, Anell B, et al. Patients with high-disease-activity relapsing-remitting multiple sclerosis in real-world clinical practice: A population-based study in Sweden. *Clin Ther* 2019; 42: 240–250.
16. Karim ME, Gustafson P, Petkau J, et al. Marginal structural Cox models for estimating the association between β-interferon exposure and disease progression in a multiple sclerosis cohort. *Am J Epidemiol* 2014; 180: 160–171.
17. Ziemssen T, Hillert J and Butzkueven H. The importance of collecting structured clinical information on multiple sclerosis. *BMC Med* 2016; 14: 81.
18. Desai RJ and Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: A primer for practitioners. *BMJ* 2019; 367: l5657.
19. Desai RJ, Rothman KJ, Bateman BT, et al. A propensity score based fine stratification approach for confounding adjustment when exposure is infrequent. *Epidemiology* 2017; 28(2): 249–257.
20. Austin PC. A tutorial and case study in propensity score analysis: An application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate Behav Res* 2011; 46(1): 119–151.