Ψ_Σ

Att skriva Rasch-baserade manus

Peter Hagell, RN PhD

The PRO-CARE Group, Faculty of Health Sciences, Kristianstad University, Kristianstad, Sweden

ERRTG
European Rasch Research and Teaching Group

Journal of Applied Measurement

GUIDELINES FOR MANUSCRIPTS

Reprinted from Smith, R.M., Linacre, J.M., and Smith, Jr., E.V. (2003). Guidelines for Manuscripts. *Journal of Applied Measurement, 4*, 198-204.

Included in this editorial are guidelines for manuscripts submitted to the ***Journal of Applied Measurement*** that involve applications of Rasch measurement. These guidelines may also be of use to those attempting to publish Rasch measurement applications in other journals that may not be familiar with these methods.

Following the guidelines, we provide a list of references that may assist individuals in gaining an overview of some of the material discussed in the guidelines. The guidelines and the list of references are by no means exhaustive. If you feel an important reference has been left out or have a recommendation for the guidelines, please e-mail us your suggestions (rsmith@jampress.org, mike@winsteps.com, or evsmith@uic.edu).

Finally, we consider this a work in progress and thank William Fisher and George Karabatsos for comments on an earlier version. We will attempt to incorporate ideas and references as we receive them. Please periodically visit the journal website at http://www.jampress.org for the most recent updates.

A. Describing the problem

1. Adequate references, at least reference to Georg Rasch (1960) when appropriate.
2. Adequate theory, at least exact algebraic representation of the Rasch model(s) used and citation for primary developer(s).
3. Adequate description of the measurement problem, including hypothesized definition of latent variable, identification of facets under investigation, description of rating scales or response formats.
4. Rationale for using Rasch measurement techniques. For example, this may include the preferen
Rasch models embody, the goal of establishing generalized reference standard metrics, or empiric
example, a comparison of the generalizability of the estimated parameters obtained from competi
rationale for using Rasch measurement is particular important when reviewers are more familiar
Response Theory or True Score Theory.

B. Describing the analysis

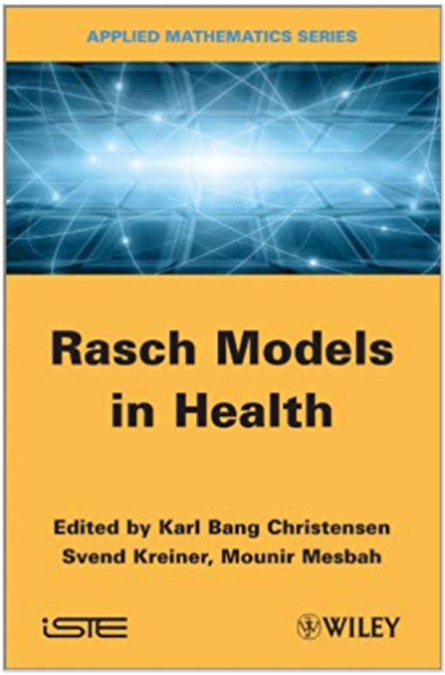
1. Name and citation or adequate description of software or estimation methodology employed.
2. Provide a rationale for the choice of fit statistics and the criteria employed to indicate adequate
This should include some acknowledgment of the Type I error rate that the critical values imply.
symmetric statistic. A value of 0.7 is further from 1.0 than is 1.3. Using a 1.3/0.7 cutoff for mean
error rate for the upper and lower tail of the mean square distribution.

Arthritis & Rheumatism (Arthritis Care & Research)
Vol. 57, No. 8, December 15, 2007, pp 1358–1362
DOI 10.1002/art.23108

SPECIAL ARTICLE

The Rasch Measurement Model in Rheumatology: What Is It and Why Use It? When Should It Be Applied, and What Should One Look for in a Rasch Paper?

ALAN TENNANT[1] AND PHILIP G. CONAGHAN[2]



Chapter 19

Reporting a Rasch Analysis

19.1. Introduction

19.1.1. ***Objectives***

The Rasch model [RAS 60] is based on a philosophy of measurement that differs considerably from the predominant paradigm in the social sciences, understood as a collective term for sciences outside the domain of the natural sciences. Therefore, reporting a Rasch analysis raises questions as to what to include in a write-up. This chapter suggests a structure of a Rasch analysis report and its elements that corresponds to the philosophy of Rasch measurement. Although it is primarily meant for writing a manuscript, the concept also helps interpret and evaluate Rasch reports. The proposed scheme is on no account an empirical compilation of what sort of findings published Rasch papers actually report. Such an account would merely ensure compliance with what has been done up to now. By contrast, the orientation of this chapter is forward-looking. Furthermore, it is in no way implied that papers deviating from the suggested scheme are necessarily deficient. Conversely, there is no guarantee that papers that do follow the proposal are based on an adequate Rasch analysis. Throughout this chapter it is assumed that the instrument development and the underlying Rasch analysis are appropriate. Wolfe and Smith [WOL 07a] provide a good overview of proper instrument development, while Wolfe and Smith [WOL 07b] refer to measure validation using Rasch models. Tennant and Conaghan [TEN 07] focus on the fields of applications of the Rasch model in health and provide guidelines as to what to look for in a Rasch report. Hagquist *et al*. [HAG 09] demonstrate the potential of Rasch analysis in nursing research [HAG 09].

Chapter written by Thomas SALZBERGER.

Introduction

No single scheme meets the requirements of every Rasch analysis under all circumstances, e.g.,

- Purpose of analysis, e.g.,
  - Existing scale vs. scale development
  - Applied vs. methodological research
- Scientific dicipline, tradition and familiarity with RMT
- Journal restrictions

Introduction

- The ambitious goals of a Rasch measurement analysis can only be fully met, if the substantive theory of the latent variable is sophisticated enough to not only suggest suitable items but also propose at least a theory-driven order of the items
- The measurement of a quantitative latent variable always refers to a frame of reference, within which specific objectivity holds. Thus, the frame of reference is defined by the conditions under which comparisons are invariant… It goes without saying that a wide frame of reference is desirable.

Introduction

- In a real data set, deviations from perfection as prescribed by the measurement model occur almost inevitably.
- Reasons for misfit have to be disentangled, which can be an intricate task.
- As a rule, changes to the data set, for example the deletion of an item, rescoring of the responses or splitting an item, should be kept to a minimum at each stage of the analysis.
- Theoretical considerations have to precede data analysis. If empirical findings inform the theory, the research becomes exploratory and should therefore be marked as such.

Suggested elements

Construct definition and operationalization
Response format and scoring

Element	Report
Latent variable Operationalization	Definition and substantive theory of the latent variable Description of the instrument (items) based on the definition of the construct, construct map
Response format and scoring	Characteristics of response scale (scoring key, number of categories, direction, position in the instrument, verbal labeling or description, etc.)

Comment:
A substantive theory suitable for RMT goes beyond a merely qualitative description; it allows for a testable hypothesis of the structure of the construct.

Sample and sampling design

Element	Report
Total population	Definition of total population as a part of the intended frame of reference sampling frame (from which the sample has actually been drawn)
Intended sample	Design, sampling method sample size

Comment:
No distributional assumtions, but consider frame of reference and targeting.

Data

Element	Report
Actual sample	Actual sample size Targeting (discuss potential problems due to poor targeting) Missing values (frequency, type and consequences) Sample characteristics (demographic variables)
Data dependency	Structure of the data in terms of dependency (e.g. repeated measurement) Consequences
Context factors	Context factors conditions and circumstances under which data were collected, to be considered when interpreting outcome of analysis

Measurement model and technical aspects

Element	Report
Fundamentals of Rasch measurement	Fundamental elements of the Rasch model (model parameters and their meaning) Unique advantages of the Rasch model/Rasch measurement theory At least stress invariance property/specific objectivity as a requirement of measurement
Measurement model	Variant of model used (depending on given data/response format), cite relevant references (e.g. [AND 78a, AND 78b, MAS 82] for polytomous Rasch model)

Comment:
The extent to which fundamental properties of Rasch measurement should be explained depends on the target audience, the degree of acceptance of the model in the field of research and on the available space.

Measurement model and technical aspects

Element	Report
Estimation method	Estimation method used (e.g. conditional maximum likelihood (CML), marginal maximum likelihood (MML) for item parameter estimation), maximum likelihood (ML) or weighted maximum likelihood (WML) for person location estimation) Often a consequence of software chosen, be aware of theoretical consequences Provide references (e.g. [MOL 95, AND 03, ZWI 95, WAR 89])
Software	Software used for data analysis, provide reference

Fit analysis

Element	Report
Local independence	Method used to investigate local independence, extent of actual local dependence in the data, plausible explanation of why local dependence occurs, remedies undertaken (e.g. item removal and sub tests)
Unidimensionality	Method used to check for unidimensionality, extent of departure from unidimensionality and remedies undertaken to resolve multidimensionality
Functioning of response scale	Indicate order of threshold estimates and any problems with empirical threshold order, ideally along with plausible interpretation of why disordering occurs, report collapsing and new scoring scheme
Invariance	Method used to check independence of item (person) parameters from respondents (items), see also DIF

Fit analysis

Comment:
The evaluation of a measurement instrument should not be made purely on the basis of statistical evidence. Numerical results need to be accompanied by qualitative interpretation and theoretical considerations.
… anomalies in the data should not simply be accounted for but revealed and exposed. Attempts at plausible explanations are certainly advantageous for future revisions of the scale, should the necessity arise.

Item fit assessment

Element	Report
Test of total fit	Type of test statistic(s) used, (e.g. χ^2 (item–trait interaction), interpretation based on sample size, targeting and person separation. Discuss theoretical implications of item deletion
Test of individual item fit	Type of test statistics used, complement fit analysis by investigating graphics
Differential item functioning (DIF)	Method used to assess DIF, measures undertaken to account for DIF, implications for substantive theory of construct and the frame of reference

Comment:
Results of all tests of fit have to be put into perspective, in particular regarding power issues and implications of misfit or marginal fit.

Person fit assessment

Element	Report
Test of individual person fit	Type of test statistic used, number or proportion of respondents outside acceptable limits Considerations of factors responsible for person misfit if persons are deleted, provide a rationale and a description of discarded respondents

Comment:
Person misfit implies that the measurement instrument does not work for some respondents as it does for most others.
If there are systematic patterns of person misfit, group means may be seriously distorted and mean comparisons invalidated.

Information

Element	Report
Targeting	Present targeting plot and/or verbal description, refer to purpose of scale when interpreting targeting, discuss consequences for person separation and power of the tests of fit
Precision	Provide estimate of person separation and standard errors at critical levels of the latent variable (e.g. at cutoff values considered important from a clinical perspective), discuss possible reasons for a low person separation index
Power of test of fit	Comment on limitations of the power of the test of fit due to, for example, targeting or sample size

Validity

Element	Report
Fit of data to the model	*See measurement requirements and fit assessment*
Matching substantive theory of latent variable and empirical evidence	Compare actual item hierarchy with expected hierarchy based on substantive theory of the construct
Comparison with other instruments	Compare instrument with other scales measuring the same latent variable, findings related to deriving a common metric by linking existing instruments (if applicable)

Comment:
Whether the scale represents a valid and generalizable instrument depends on the degree to which the analysis was confirmatory. If a large set of items has been reduced to a relatively small subset and/or the data have been altered extensively (e.g. by rescoring or item splitting), we run the risk of capitalizing on chance.

Application and usefulness

Element	Report
Description of final instrument	List set of items in the final scale (after deletion of misfitting items) and, in case of rescoring, the final scoring scheme
Item parameters	Table of final threshold estimates and overall item locations
Person parameters	State person summary statistics (mean, SD) and describe shape of distribution
Theoretical implications	Reflect the consequences of scale purification (in particular item deletion) for theory of the construct
Application and usefulness of the scale	Findings related to the application of the scale and its relationship to other constructs

Application and usefulness

Element	Report
Recommendations	Provide recommendations for scale usage, stress strong and weak points of the instrument, suggest amending of scoring procedures in future applications when appropriate (e.g. decrease number of categories in case of disordered thresholds, increase number of categories if thresholds are properly ordered but precision is too low), propose changes to items that do not function properly

Comment:
Traditionally, the relationship of the latent variable and other constructs is integrated into the concept of validity under the label of external validity. However, the assessment of external relationships provides at best indirect evidence of validity. External relationships should be better viewed as aspects of a scale's usefulness.

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 1–19: Editorial
https://doi.org/10.1080/00031305.2019.1583913

Taylor & Francis
Taylor & Francis Group

EDITORIAL

OPEN ACCESS

Check for updates

Moving to a World Beyond "$p < 0.05$"

Some of you exploring this special issue of *The American Statistician* might be wondering if it's a scolding from pedantic statisticians lecturing you about what *not* to do with *p*-values, without offering any real ideas of what *to do* about the very hard problem of separating signal from noise in data and making decisions under uncertainty. Fear not. In this issue, thanks to 43 innovative and thought-provoking papers from forward-looking statisticians, help is on the way.

1. "Don't" Is Not Enough

There's not much we can say here about the perils of *p*-values and significance testing that hasn't been said already for decades (Ziliak and McCloskey 2008; Hubbard 2016). If you're just arriving to the debate, here's a sampling of what not to do:

- Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" (i.e., the *p*-value passed some arbitrary threshold such as $p < 0.05$).
- Don't believe that an association or effect exists just because it was statistically significant.
- Don't believe that an association or effect is absent just because it was not statistically significant.
- Don't believe that your *p*-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

Don't. Don't. Just…don't. Yes, we talk a lot about don'ts. The *ASA Statement on p-Values and Statistical Significance* (Wasserstein and Lazar 2016) was developed primarily because after decades, warnings about the don'ts had gone mostly unheeded. The statement was about what not to do, because there is widespread agreement about the don'ts.

Knowing what not to do with *p*-values is indeed necessary, but it does not suffice. It is as though statisticians were asking users of statistics to tear out the beams and struts holding up the edifice of modern scientific research without offering solid construction materials to replace them. Pointing out old, rotting timbers was a good start, but now we need more.

Recognizing this, in October 2017, the American Statistical Association (ASA) held the Symposium on Statistical Inference, a two-day gathering that laid the foundations for this special issue of *The American Statistician*. Authors were explicitly instructed to develop papers for the variety of audiences interested in these topics. If you use statistics in research, business, or policymaking but are not a statistician, these articles were indeed written with YOU in mind. And if you are a statistician, there is still much here for you as well.

The papers in this issue propose many new ideas, ideas that in our determination as editors merited publication to enable broader consideration and debate. The ideas in this editorial are likewise open to debate. They are our own attempt to distill the wisdom of the many voices in this issue into an essence of good statistical practice as we currently see it: some do's for teaching, doing research, and informing decisions.

Yet the voices in the 43 papers in this issue do not sing as one. At times in this editorial and the papers you'll hear deep dissonance, the echoes of "statistics wars" still simmering today (Mayo 2018). At other times you'll hear melodies wrapping in a rich counterpoint that may herald an increasingly harmonious new era of statistics. To us, these are all the sounds of statistical inference in the 21st century, the sounds of a world learning to venture beyond "$p < 0.05$."

This is a world where researchers are free to treat "$p = 0.051$" and "$p = 0.049$" as not being categorically different, where authors no longer find themselves constrained to selectively publish their results based on a single magic number. In this world, where studies with "$p < 0.05$" and studies with "$p > 0.05$" are not automatically in conflict, researchers will see their results more easily replicated—and, even when not, they will better understand *why*. As we venture down this path, we will begin to see fewer false alarms, fewer overlooked discoveries, and the development of more customized statistical strategies. Researchers will be free to communicate all their findings in all their glorious uncertainty, knowing their work is to be judged by the quality and effective communication of their science, and not by their *p*-values. As "statistical significance" is used less, statistical thinking will be used more.

The *ASA Statement on P-Values and Statistical Significance* started moving us toward this world. As of the date of publication of this special issue, the statement has been viewed over 294,000 times and cited over 1700 times—an average of about 11 citations per week since its release. Now we must go further. That's what this special issue of *The American Statistician* sets out to do.

To get to the do's, though, we must begin with one more don't.

Work on new reporting guidelines initiated by the ERRTG

ERRTG

European
Rasch
Research and
Teaching
Group