PLYMOUTH UNIVERSITY PENINSULA SCHOOLS OF MEDICINE & DENTISTRY

# Approaches to advancing patient focussed outcomes assessment in clinical trials

Jeremy Hobart MD, PhD

Consultant Neurologist
Professor of Clinical Neurology and Health Measurement

EMA workshop on Multiple Sclerosis, London 17TH October 2013

# Workshop goals

**Objectives of the planned workshop**

In response to the considerable interest created by the on-going revision of the current MS guideline, the EMA has decided to provide an opportunity for the different stakeholders to come together and discuss the key scientific issues in the field. The main goal of the workshop is to make sure that in the revision of the MS guideline EMA can take into consideration the most up-to-date, state of the art scientific developments in multiple sclerosis, as well as the positions of the experts in the field on the main topics in the guideline.

# Overview

- EMA draft guidance document – critique
- Other guidance documents - relevance
- Evidence / advances in measurement science

# 5. Criteria for assessment of efficacy in confirmatory trials

## *5.1. Treatments for acute relapses*

Duration and severity of relapses and overall recovery or prevention of their sequelae are relevant parameters.

If, for a test drug an effect on the duration, severity and/or recovery from a relapse is claimed, this claim should be based on clinical trials with methylprednisolone as a positive control and a placebo arm for the internal validation of the study. Such study should include early escape conditions to allow rescue treatment when the patient fails to improve or worsens. Patients should be followed for an appropriate time (e.g. at least 6 months) after each relapse to be sure that the degree of recovery after the relapse is well assessed.

Alternative study designs may be a superiority trial versus methylprednisolone, or a placebo controlled trial in the add-on setting i.e. on top of corticosteroids. As there is no consensus concerning the corticosteroid dosage regimen in context of a clinical trial, the corticosteroid regimen should be standardized.

The impact of those acute treatments on the subsequent course of the disease (rate and severity of further relapses, progression of disability, even change from relapsing remitting into SPMS) is also relevant.

## *5.2. Treatments aiming to modify the natural course of the disease*

### 5.2.1. Primary efficacy parameters

A distinction should be made between accumulation of disability in relation to relapses in RRMS and progression of disability in SPMS or in PPMS.

The primary efficacy parameter in confirmatory trials in SPMS and in PPMS should be a clinically measured prevention or delay of the disability progression.

In patients with RRMS or SPMS with superimposed relapses (RMS), the primary efficacy parameter may be the relapse rate although it cannot be taken as a surrogate for disease progression and this would be expressed accordingly in the SmPC. Moreover, progression of disability should be evaluated and worsening of disability should be reasonably excluded by means of adequately powered long-term studies.

It would be highly desirable also to evaluate if the effect on progression is maintained on a long-term basis.

### 5.2.2. Secondary efficacy endpoints

- Disability. In studies where it is not the primary variable, it is a very important secondary endpoint that should be evaluated.
- Relapses. Recommended parameters are the rate of relapses (in studies where it is not the primary efficacy parameter), frequency of moderate/severe relapses, proportion of patients free from relapses at a given time, time to first relapse, proportion of subjects receiving rescue therapy, number of relapses.
- MRI derived parameters.
- Absence of disease activity i.e. absence of relapse and MRI-activity
- Other measures related to progression of disability supplementary to the measure chosen such as the primary variable (e.g. neurological rating scales, measures of cognitive impairment, fatigue scales, ambulatory index.

# 6. Methods to assess efficacy

## *6.1. Progression of disability*

The Kurtz's Expanded Disability Status Scale (EDSS) is the most widely used and known scale to assess changes in disability in MS.

The disadvantages and advantages of the EDSS in assessing disability in MS are well known. Therefore, on the one hand, the development of alternative scales for assessing disability in MS is advocated since these scales, if validated and justified, may be more appropriate than the EDSS. On the other hand, the EDSS should still be used in order to facilitate comparisons with other studies.

As the EDSS has a limited inter and intra-observer reliability, all possible actions intended to increase reliability of the scale should be adopted: training of observers, same physician evaluating the patient throughout the trial, standardised times and schedules for assessments, standardised protocols for neurological examination, measured distances for assessments of mobility and definitions of all the terms used. The mean change in score from the baseline is not an appropriate efficacy parameter. Based on EDSS scores, treatment failure or progression should be predefined e.g. as the achievement of a specified degree of disability or of a sustained worsening of relevant magnitude (1 point when EDSS scores $\leq$ 5.5; 0.5 points if baseline score is > 5.5). Acceptable efficacy parameters endpoints are the time to reach progression or the proportion of individuals who have shown progression at a pre-specified time.

Accurate and reliable definition of sustained worsening is important and should include two consecutive examinations carried out by the same physician at least 6 months apart.

As a supportive parameter, disability can also be expressed by summary measures obtained from serial measures at scheduled visits, indicating the degree of disability experienced by the patient during a period of time, disregarding whether it is in relation to relapses or not. It is recognised that the EDSS does not adequately assess upper limb function and cognitive impairment and the use of specific methods could be useful. In this context, additional neurological rating scales, quantitative neuron-performance tests (e.g. MSFC) or patient and neurologist global opinion may be used as secondary measurements of disability.

## *6.2. Relapses*

The annualised relapse rate is an acceptable parameter to assess relapses. The definition a priori of responders in terms of absence of relapses is recommended.

Identification of a relapse may be difficult as patients frequently suffer from pseudo-exacerbations caused by infection, heat, or stress. An accurate definition of relapse (their occurrence, time of beginning, time of ending, minimum duration to qualify as a relapse, maximum time elapsed between two symptoms to qualify as a single relapse, severity) should be included in clinical trials. Identification of relapses should be blinded to therapy. The use of corticosteroids (or other concomitant therapies) for the treatment of acute relapses that may occur throughout the trial should be carefully standardised.

Even if an effect on relapses may be shown within one year, a maintained effect on relapses should be demonstrated at least during two years. Time to next (second relapse) is not considered a good efficacy parameter.

The analysis model should be specified in the study protocol and ensure type-1 error is controlled including reasonable assumptions regarding the variance. Furthermore, the impact of premature withdrawal needs to be explored based on reasonable assumptions of the expected relapse rate in the missing observation time. A sensitivity analysis is recommended. Reference is made to the CHMP guideline on missing data (see section 3).

## *6.4. Quality of Life (QoL)*

Few data are available on validation of specific instruments for QoL in patients suffering MS. If evaluation of QoL in MS is considered, reliable and validated scales should be used. Results, if considered relevant, may be mentioned in section 5.1 of the SmPC.

# Critique

- Highlights importance of COA's
- Limited specifics on:
  - COA selection
  - EMA expectations of selected COAs
  - Measurement (psychometric) method

# Relevance of other regulatory guidance

- FDA also highlights importance of COA's

- More detailed specifics:
  - COA selection
  - COA expectations
  - Methods

# Roadmap to **PATIENT-FOCUSED OUTCOME MEASUREMENT** in Clinical Trials

## 1 Understanding the Disease or Condition

**A. Natural history of the disease or condition**
- Onset/Duration/Resolution
- Diagnosis
- Pathophysiology
- Range of manifestations

**B. Patient subpopulations**
- By severity
- By onset
- By comorbidities
- By phenotype

**C. Health care environment**
- Treatment alternatives
- Clinical care standards
- Health care system perspective

**D. Patient/caregiver perspectives**
- Definition of treatment benefit
- Benefit-risk tradeoffs
- Impact of disease

## 2 Conceptualizing Treatment Benefit

**A. Identify concept(s) of interest (COI) for meaningful treatment benefit, i.e., How a patient:**
- Survives
- Feels (e.g., symptoms)
- Functions

**B. Define context of use (COU) for clinical trial:**
- Disease/Condition entry criteria
- Clinical trial design
- Endpoint positioning

**C. Select clinical outcome assessment (COA) type:**
- Patient-Reported Outcome (PRO)
- Observer-Reported Outcome (ObsRO)
- Clinician-Reported Outcome (ClinRO)
- Performance Outcome (motor, sensory, cognition)

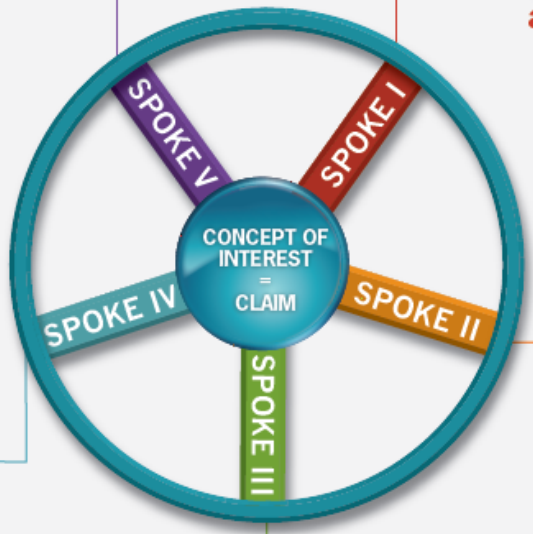
## 3 Selecting/Developing the Outcome Measure

**A. Search for existing COA measuring COI in COU:**
- Measure exists
- Measure exists but needs to be modified
- No measure exists
- Measure under development

**B. Begin COA development**
- Document content validity (qualitative or mixed methods research)
- Evaluate cross-sectional measurement properties (reliability and construct validity)
- Create user manual
- Consider submitting to FDA for COA qualification as exploratory endpoint

**C. Complete COA development:**
- Document longitudinal measurement properties (construct validity, ability to detect change)
- Document guidelines for interpretation of treatment benefit and relationship to claim
- Update user manual
- Submit to FDA for COA qualification as effectiveness endpoint to support claims

# Qualification of **CLINICAL OUTCOME ASSESSMENTS** (COAs)




## I. Identify Context of Use (COU) and Concept of Interest (COI)

- Outline hypothesized concepts and potential claims
- Determine intended population
- Determine intended application/characteristics (type of scores, mode and frequency of administration)
- Perform literature/expert review
- Develop hypothesized conceptual framework
- Position COA within a preliminary endpoint model
- Document COU and COI

## II. Draft Instrument and Evaluate Content Validity

- Obtain patient or other reporter input
- Generate new items
- Select recall period, response options and format
- Select mode/method of administration/data collection
- Conduct cognitive interviewing
- Pilot test draft instrument
- Finalize instrument content, format and scoring rule
- Document content validity

## III. Cross-sectional Evaluation of Other Measurement Properties

- Assess score reliability (test-retest or inter-rater) and construct validity
- Establish administration procedures & training materials
- Document measure development
- Prepare user manual
- Consider submitting to FDA for COA qualification as exploratory endpoint prior to longitudinal evaluation

## IV. Longitudinal Evaluation of Measurement Properties/ Interpretation Methods

- Assess ability to detect change and construct validity
- Identify responder definition(s)
- Provide guidelines for interpretation of treatment benefit and relationship to claim
- Document all results
- Update user manual
- Submit to FDA for COA qualification as effectiveness endpoint to support claims

## V. Modify Instrument

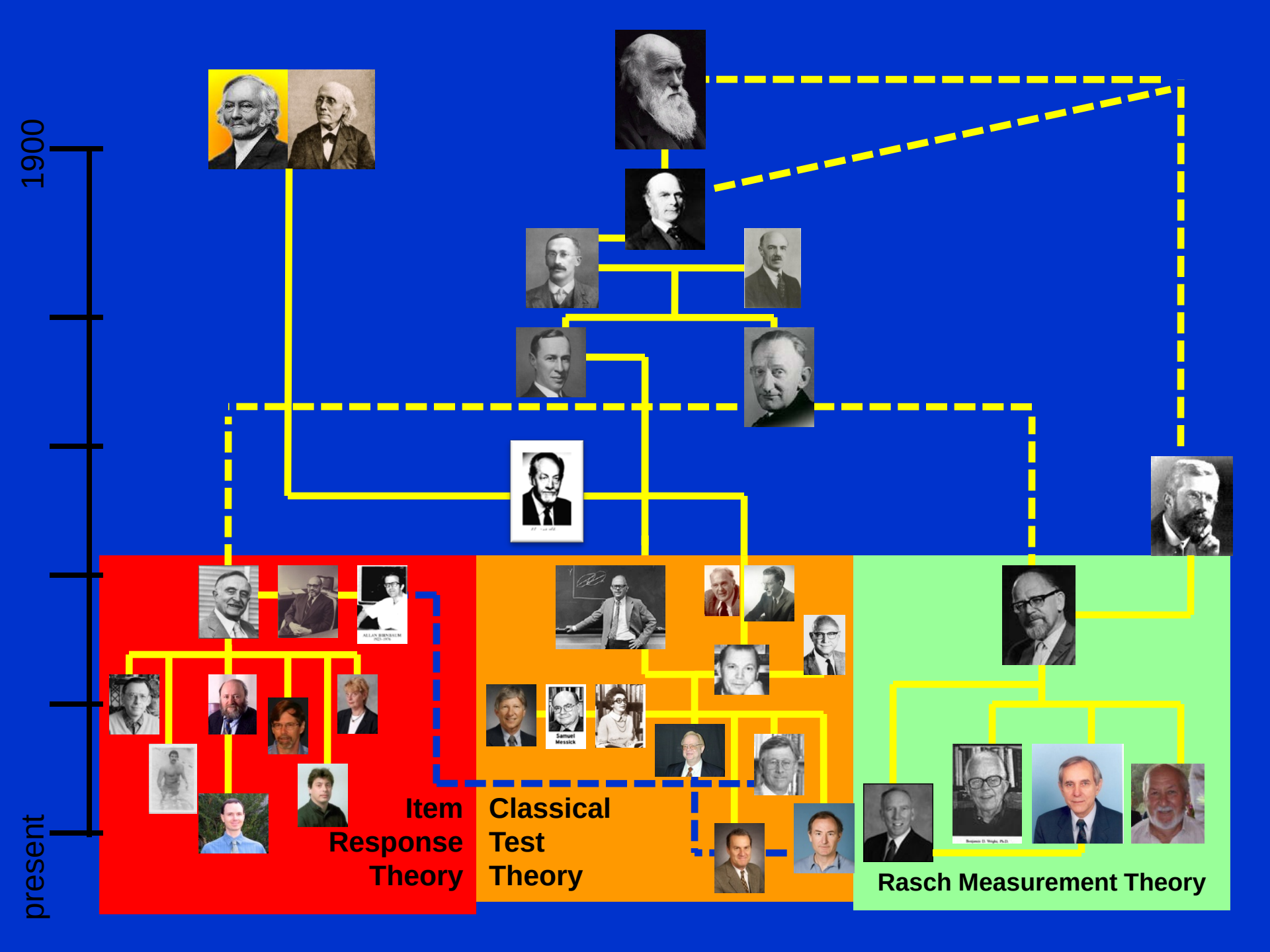
- Identify a new COU
- Change wording of items, response options, recall period, or mode/method of administration/data collection
- Translate and culturally adapt
- Evaluate modifications using spokes I - IV
- Document all changes
- Consider submitting to FDA for qualification of new COA, as appropriate

U.S. Food and Drug Administration
Center for Drug Evaluation and Research
Office of New Drugs
http://www.fda.gov/Drugs

# Advances from measurement science

- Application of new psychometric methods
  - COA evaluation
  - COA modification
  - COA development

- Conceptual clarity


1900
present
Item Response Theory
Classical Test Theory
Rasch Measurement Theory
Samuel Messick

# What do new methods tell us?

# “Legacy” measures have critical limitations….that can be improved / fixed

ELSEVIER

Alzheimer's & Dementia 9 (2013) S10–S20

Alzheimer's & Dementia

## Putting the Alzheimer's cognitive test to the test II: Rasch Measurement Theory

Jeremy Hobart[a,*,†], Stefan Cano[a,†], Holly Posner[b], Ola Selnes[c], Yaakov Stern[d], Ronald Thomas[e], John Zajicek[a]; for the Alzheimer's Disease Neuroimaging Initiative[‡]

[a]*Clinical Neurology Research Group, Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth, UK*
[b]*Clinical R&D, Pfizer, Inc., New York, NY, USA*
[c]*Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[d]*Cognitive Neuroscience Division, Columbia University, New York, NY, USA*
[e]*Department of Family and Preventative Medicine, University of California, San Diego, CA, USA*

**Abstract**

**Background:** The Alzheimer's Disease Assessment Scale—Cognitive Behavior section (ADAS-Cog) is the most widely used measure of cognitive performance in AD clinical trials. This key role has rightly brought its performance under increased scrutiny with recent research using traditional psychometric methods, questioning the ADAS-Cog's ability to adequately measure early-stage disease. However, given the limitations of traditional psychometric approaches, herein we use the more sophisticated Rasch Measurement Theory (RMT) methods to fully examine the strengths and weaknesses of the ADAS-Cog, and identify potential paths toward its improvement.

**Methods:** We analyzed AD Neuroimaging Initiative (ADNI) ADAS-Cog data (675 measurements across four time-points over 2 years) from the AD participants. RMT analysis was undertaken to examine three broad areas: adequacy of scale-to-sample targeting; degree to which, taken together, the ADAS-Cog items adequately perform as a measuring instrument; and how well the scale measured the subjects in the current sample.

**Results:** The 11 ADAS-Cog components mapped-out a measurement continuum, worked together adequately, and were stable across different time-points and samples. However, the scale did not prove to be a good match to the patient sample supporting previous research. RMT analysis also identified problematic “gaps” and “bunching” of the components across the continuum.

**Conclusion:** Although the ADAS-Cog has the building blocks of a good measurement instrument, this sophisticated analysis confirms limitations with potentially serious implications for clinical trials. Importantly, and unlike traditional psychometric methods, our RMT analysis has provided important clues aimed at solving the measurement problems of the ADAS-Cog.



*Keywords:* Alzheimer's disease; Clinical trials; Psychometrics; Reliability; Validity; Rasch Measurement Theory

# Clinically meaningful paths to scale improvement are clear….

*Research Paper*

MULTIPLE SCLEROSIS JOURNAL | MSJ

*Multiple Sclerosis Journal*
0(0) 1–10

DOI: 10.1177/1352458512462919
msj.sagepub.com
SAGE

## Can the ABILHAND handle manual ability in MS?

**LE Barrett, SJ Cano, JP Zajicek and JC Hobart**

**Abstract**

**Background:** Hand dysfunction is common in multiple sclerosis (MS). Recent interest has focused on incorporating patient-reported outcome (PRO) instruments into clinical trials. Nevertheless, examinations are rare in MS of existing manual ability measures.

**Objectives:** The objective of this paper is to evaluate the 23-item ABILHAND, developed for use after stroke, in people with MS, comparing the findings from two psychometric approaches.

**Methods:** We analysed ABILHAND data from 300 people with MS using: 1) traditional psychometric methods (data completeness, scaling assumptions, reliability, internal and external construct validity); and 2) Rasch measurement methods (including targeting, item response category ordering, data fit to the Rasch model, spread of item locations, item scoring bias, item stability, reliability, person response validity).

**Results:** Traditional psychometric methods implied ABILHAND was reliable and valid in this sample. Rasch measurement methods supported this finding. The three-category scoring function worked as intended and item fit to Rasch model expectations was acceptable. The 23 items (location range −3.16 to +2.73 logits) mapped a continuum of manual ability. Reliability was high (Person Separation Index (PSI) = 0.95).

**Conclusion:** Both psychometric evaluations supported ABILHAND as a robust manual ability PRO measure for MS. Rasch measurement methods were more informative and, consistent with its role of detecting anomalies, identified ways of advancing further ABILHAND's measurement performance to reduce any potential for type II errors in clinical trials.

# Standard analysis methods underestimate changes and difference…



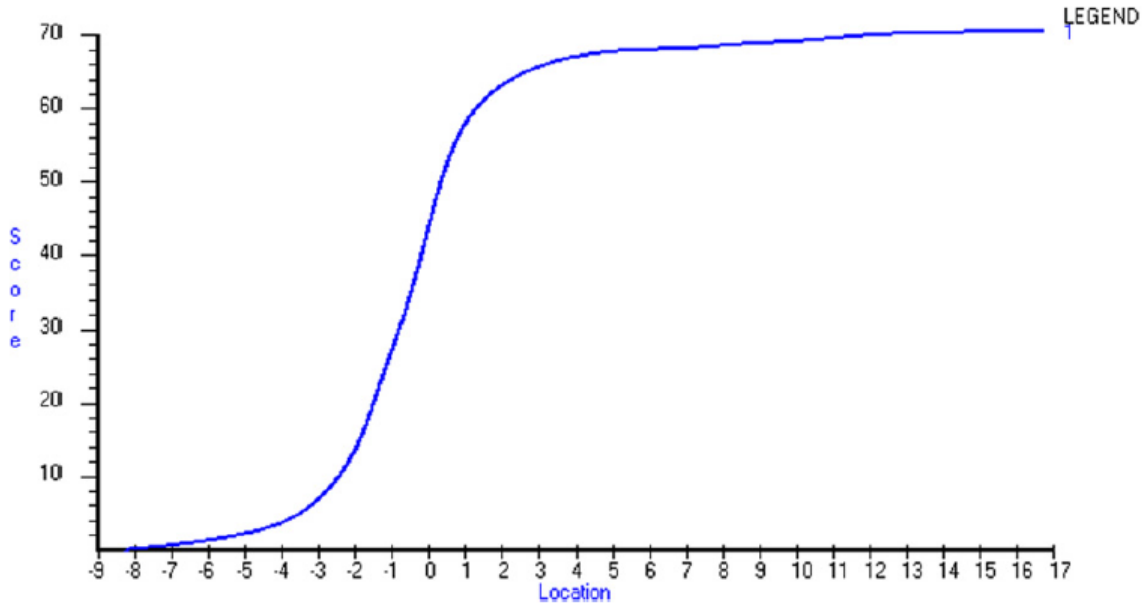



Fig. 4. ADAS-Cog raw score to interval level across the sample. The figure shows the relationship between ADAS-Cog total scores (which are ordinal and therefore have an unequal interval), and the linear measurement they imply (which are equal intervals) is S-shaped. The change in cognitive performance implied by a change of 1 point in ADAS-Cog total score varies eightfold across the subscale range.

# Traditional methods of evaluation change, and comparing ability to detect change, are misleading

Research paper

## Effect sizes can be misleading: is it time to change the way we measure change?

Jeremy C Hobart,[1,2] Stefan J Cano,[1,2] Alan J Thompson[2]

**See Editorial Commentary, p 943**

► Additional appendix is published online only. To view these files please visit the journal online (http://jnnp.bmj.com).

[1]The Neurological Outcome Measures Unit, Peninsula College of Medicine and Dentistry, Devon, UK
[2]UCL Institute of Neurology, London UK

**Correspondence to**
Dr J C Hobart, Department of Clinical Neuroscience, Peninsula College of Medicine and Dentistry, Room N16 ITTC Building, Tamar Science Park, Davy Road, Plymouth, Devon PL6 8BX, UK;
jeremy.hobart@pms.ac.uk

Received 24 November 2009
Revised 6 April 2010
Accepted 13 April 2010
Published Online First
13 July 2010

**ABSTRACT**

**Objectives** Previous comparisons of the ability to detect change in the Barthel Index (BI) and Functional Independence Measure motor scale (FIMm) have implied these two scales are equally responsive when examined using traditional effect size statistics. Clinically, this is counterintuitive as the FIMm has greater potential to detect change than the BI and raises concerns about the validity of effect size statistics as indicators of rating scale responsiveness. To examine these concerns, in this study a sophisticated psychometric analysis was applied, Rasch measurement to BI and FIMm data.

**Methods** BI and FIMm data were examined from 976 people at a single neurorehabilitation unit. Rasch analysis was used to compare the responsiveness of the BI and FIMm at the group comparison level (effect sizes, relative efficiency, relative precision) and for each individual person in the sample by computing the significance of their change.

**Results** *Group level* analyses from both interval measurements and ordinal scores implied the BI and FIMm had equivalent responsiveness (BI and FIMm effect size ranges −0.82 to −1.12 and −0.77 to −1.05, respectively). However, *individual person level* analyses indicated that the FIMm detected significant improvement in almost twice as many people as the BI (50%, n=496 vs 31%, n=298), and recorded less people as unchanged on discharge (FIMm=4%, n=38; BI=12%, n=115). This difference was found to be statistically significant ($\chi^2$=273.81; p<0.000).

**Conclusions** These findings demonstrate that effect size calculations are limited and potentially misleading indicators of rating scale responsiveness at the group comparison level. Rasch analysis at the individual person level showed the superior responsiveness of the FIMm, supporting clinical expectation, and its added value as a method for examining and comparing rating scale responsiveness.

more response categories. We tested this hypothesis by comparing the BI with the FIMm, a scale that uses the same items but has more item response categories.[9] Results showed that the FIMm had greater *potential* to detect change (smaller item and total score floor and ceiling effects than the BI) and detected change in more people undergoing rehabilitation. Despite this evidence of better *potential* to detect change, the FIMm and BI had almost identical effect size calculations implying the same *ability* to detect change at the group comparison level. This finding is counterintuitive clinically and questions the validity of effect size statistics as indicators of rating scale responsiveness.

To explore this issue, we examined the relative responsiveness of the BI and FIMm in the same dataset using a more sophisticated psychometric method, Rasch measurement.[10–12] This method advances the analysis of rating scale responsiveness in three specific ways. Firstly, Rasch analysis enables interval level (linear) measurements of activity limitation to be estimated from ordinal level BI and FIMm total scores. This is valuable because fixed changes in ordinal total scores (eg, 10 points) imply variable changes in interval level measurements across the scale range.[2 3 13] Thus analysing total scores may hide responsiveness differences between scales. Secondly, Rasch analysis enables a legitimate examination of changes in activity limitation at the *individual person level*, in addition to comparisons at the *group level*. In contrast, traditional psychometric analyses are not recommended for individual person decision making.[3 14] The third benefit is that Rasch analysis enables scales measuring the same construct, as the BI and FIMm purport, to be equated on a common metric.[2] This enables people's measurements on the

# Legitimate individual–person level analysis

Research paper

## Effect sizes can be misleading: is it time to change the way we measure change?

Jeremy C Hobart,[1,2] Stefan J Cano,[1,2] Alan J Thompson[2]

See Editorial Commentary, p 943

► Additional appendix is published online only. To view these files please visit the journal online (http://jnnp.bmj.com).

[1]The Neurological Outcome Measures Unit, Peninsula College of Medicine and Dentistry, Devon, UK
[2]UCL Institute of Neurology, London UK

**Correspondence to**
Dr J C Hobart, Department of Clinical Neuroscience, Peninsula College of Medicine and Dentistry, Room N16 ITTC Building, Tamar Science Park, Davy Road, Plymouth, Devon PL6 8BX, UK;
jeremy.hobart@pms.ac.uk

Received 24 November 2009
Revised 6 April 2010
Accepted 13 April 2010
Published Online First
13 July 2010

**ABSTRACT**

**Objectives** Previous comparisons of the ability to detect change in the Barthel Index (BI) and Functional Independence Measure motor scale (FIMm) have implied these two scales are equally responsive when examined using traditional effect size statistics. Clinically, this is counterintuitive as the FIMm has greater potential to detect change than the BI and raises concerns about the validity of effect size statistics as indicators of rating scale responsiveness. To examine these concerns, in this study a sophisticated psychometric analysis was applied, Rasch measurement to BI and FIMm data.

**Methods** BI and FIMm data were examined from 976 people at a single neurorehabilitation unit. Rasch analysis was used to compare the responsiveness of the BI and FIMm at the group comparison level (effect sizes, relative efficiency, relative precision) and for each individual person in the sample by computing the significance of their change.

**Results** *Group level* analyses from both interval measurements and ordinal scores implied the BI and FIMm had equivalent responsiveness (BI and FIMm effect size ranges −0.82 to −1.12 and −0.77 to −1.05, respectively). However, *individual person level* analyses indicated that the FIMm detected significant improvement in almost twice as many people as the BI (50%, n=496 vs 31%, n=298), and recorded less people as unchanged on discharge (FIMm=4%, n=38; BI=12%, n=115). This difference was found to be statistically significant ($\chi^2$=273.81; p<0.000).

**Conclusions** These findings demonstrate that effect size calculations are limited and potentially misleading indicators of rating scale responsiveness at the group comparison level. Rasch analysis at the individual person level showed the superior responsiveness of the FIMm, supporting clinical expectation, and its added value as a method for examining and comparing rating scale responsiveness.

more response categories. We tested this hypothesis by comparing the BI with the FIMm, a scale that uses the same items but has more item response categories.[9] Results showed that the FIMm had greater *potential* to detect change (smaller item and total score floor and ceiling effects than the BI) and detected change in more people undergoing rehabilitation. Despite this evidence of better *potential* to detect change, the FIMm and BI had almost identical effect size calculations implying the same *ability* to detect change at the group comparison level. This finding is counterintuitive clinically and questions the validity of effect size statistics as indicators of rating scale responsiveness.

To explore this issue, we examined the relative responsiveness of the BI and FIMm in the same dataset using a more sophisticated psychometric method, Rasch measurement.[10–12] This method advances the analysis of rating scale responsiveness in three specific ways. Firstly, Rasch analysis enables interval level (linear) measurements of activity limitation to be estimated from ordinal level BI and FIMm total scores. This is valuable because fixed changes in ordinal total scores (eg, 10 points) imply variable changes in interval level measurements across the scale range.[2 3 13] Thus analysing total scores may hide responsiveness differences between scales. Secondly, Rasch analysis enables a legitimate examination of changes in activity limitation at the *individual person level*, in addition to comparisons at the *group level*. In contrast, traditional psychometric analyses are not recommended for individual person decision making.[3 14] The third benefit is that Rasch analysis enables scales measuring the same construct, as the BI and FIMm purport, to be equated on a common metric.[2] This enables people's measurements on the

# Critical limitations of any statistical method

*Research Paper*

MULTIPLE SCLEROSIS JOURNAL | MSJ

*Multiple Sclerosis Journal*
0(0) 1–11

DOI: 10.1177/1352458513483378
msj.sagepub.com
SAGE

# Achieving valid patient-reported outcomes measurement: a lesson from fatigue in multiple sclerosis

**Jeremy Hobart[1], Stefan Cano[1], Rachel Baron[1], Alan Thompson[2], Steven Schwid[3*], John Zajicek[1] and David Andrich[4]**

## Abstract

**Background:** The increasing influence of patient-reported outcome (PRO) measurement instruments indicates their scrutiny has never been more crucial. Above all, PRO instruments should be valid: shown to assess what they purport to assess.

**Objectives:** To evaluate a widely used fatigue PRO instrument, highlight key issues in understanding PRO instrument validity, demonstrate limitations of those approaches and justify notable changes in the validation process.

**Methods:** A two-phase evaluation of the 40-item Fatigue Impact scale (FIS): a qualitative evaluation of content and face validity using expert opinion ($n$=30) and a modified Delphi technique; a quantitative psychometric evaluation of internal and external construct validity of data from 333 people with multiple sclerosis using traditional and modern methods.

**Results:** Qualitative evaluation did not support content or face validity of the FIS. Expert opinion agreed with the subscale placement of 23 items (58%), and classified all 40 items as being non-specific to fatigue impact. Nevertheless, standard quantitative psychometric evaluations implied, largely, FIS subscales were reliable and valid.

**Conclusions:** Standard quantitative 'psychometric' evaluations of PRO instrument validity can be misleading. Evaluation of existing PRO instruments requires both qualitative and statistical methods. Development of new PRO instruments requires stronger conceptual underpinning, clearer definitions of the substantive variables for measurement and hypothesis-testing experimental designs.

# “Disability”

J Neurol (2011) 258:1477–1488
DOI 10.1007/s00415-011-5963-7

ORIGINAL COMMUNICATION

## The development of ICF Core Sets for multiple sclerosis: results of the International Consensus Conference

Michaela Coenen · Alarcos Cieza · Jenny Freeman · Fary Khan · Deborah Miller · Andrea Weise · Jürg Kesselring · The members of the Consensus Conference

daily living and social participation [3–5]. Disability in people with MS comprises impaired body functions and structures as well as limitations in activities and restrictions in participation modified by contextual factors such as environmental and personal factors. However, up to now

M. Coenen · A. Cieza
ICF Research Branch in cooperation with the WHO Collaborating Centre for the Family of International Classifications in Germany (at DIMDI), Munich, Germany

J. Freeman
School of Health Professions, Allied Health Centre, University of Plymouth, Plymouth, UK

F. Khan
Department of Rehabilitation Medicine, Royal Melbourne Hospital, University of Melbourne, Melbourne, Australia

D. Miller
Mellen Center for Multiple Sclerosis Treatment and Research, Cleveland Clinic, Cleveland, USA

A. Weise · J. Kesselring
Department of Neurorehabilitation, Valens Rehabilitation Centre, Valens, Switzerland

ICF Core Set

### Introduction

Multiple sclerosis (MS) can follow a variety of clinical courses and is unpredictable in terms of prognosis. Individuals diagnosed with MS have to face various limitations in functioning and experience disability during the course of the disease [1, 2] having a significant impact on independence, employability, the performance of activities of daily living and social participation [3–5]. Disability in people with MS comprises impaired body functions and structures as well as limitations in activities and restrictions in participation modified by contextual factors such as environmental and personal factors. However, up to now

# Effect of dronabinol on progression in progressive multiple sclerosis (CUPID): a randomised, placebo-controlled trial

*John Zajicek, Susan Ball, David Wright, Jane Vickery, Andrew Nunn, David Miller, Mayam Gomez Cano, David McManus, Sharukh Mallik, Jeremy Hobart, on behalf of the CUPID investigator group*

## Summary

**Background** Laboratory evidence has shown that cannabinoids might have a neuroprotective action. We investigated whether oral dronabinol ($\Delta^9$-tetrahydrocannabinol) might slow the course of progressive multiple sclerosis.

**Methods** In this multicentre, parallel, randomised, double-blind, placebo-controlled study, we recruited patients aged 18–65 years with primary or secondary progressive multiple sclerosis from 27 UK neurology or rehabilitation departments. Patients were randomly assigned (2:1) to receive dronabinol or placebo for 36 months; randomisation was by stochastic minimisation, using a computer-generated randomisation sequence, balanced according to expanded disability status scale (EDSS) score, centre, and disease type. Maximum dose was 28 mg per day, titrated against bodyweight and adverse effects. Primary outcomes were EDSS score progression (masked assessor, time to progression of ≥1 point from a baseline score of 4·0–5·0 or ≥0·5 points from a baseline score of ≥5·5, confirmed after 6 months) and change from baseline in the physical impact subscale of the 29-item multiple sclerosis impact scale (MSIS-29-PHYS). All patients who received at least one dose of study drug were included in the intention-to-treat analyses. This trial is registered as an International Standard Randomised Controlled Trial (ISRCTN 62942668).

**Findings** Of the 498 patients randomly assigned to a treatment group, 329 received at least one dose of dronabinol and 164 received at least one dose of placebo (five did not receive the allocated intervention). 145 patients in the dronabinol group had EDSS score progression (0·24 first progression events per patient-year; crude rate) compared with 73 in the placebo group (0·23 first progression events per patient-year; crude rate); HR for prespecified primary analysis was 0·92 (95% CI 0·68–1·23; p=0·57). Mean yearly change in MSIS-29-PHYS score was 0·62 points (SD 3·29) in the dronabinol group versus 1·03 points (3·74) in the placebo group. Primary analysis with a multilevel model gave an estimated between-group difference (dronabinol–placebo) of –0·9 points (95% CI –2·0 to 0·2). We noted no serious safety concerns (114 [35%] patients in the dronabinol group had at least one serious adverse event, compared with 46 [28%] in the placebo group).

**Interpretation** Our results show that dronabinol has no overall effect on the progression of multiple sclerosis in the progressive phase. The findings have implications for the design of future studies of progressive multiple sclerosis, because lower than expected progression rates might have affected our ability to detect clinical change.

*Lancet Neurol* 2013; 12: 857–65

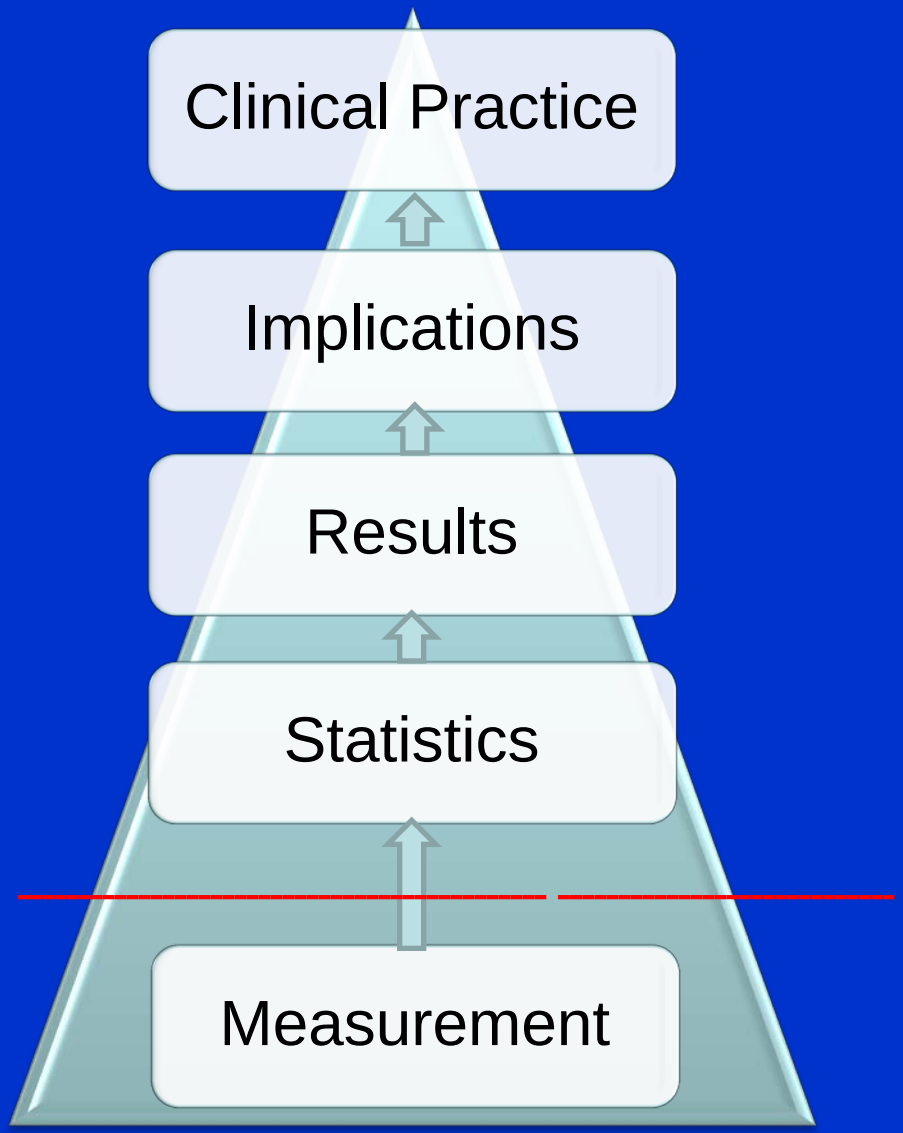
Published Online
July 13, 2013
http://dx.doi.org/10.1016/S1474-4422(13)70159-5

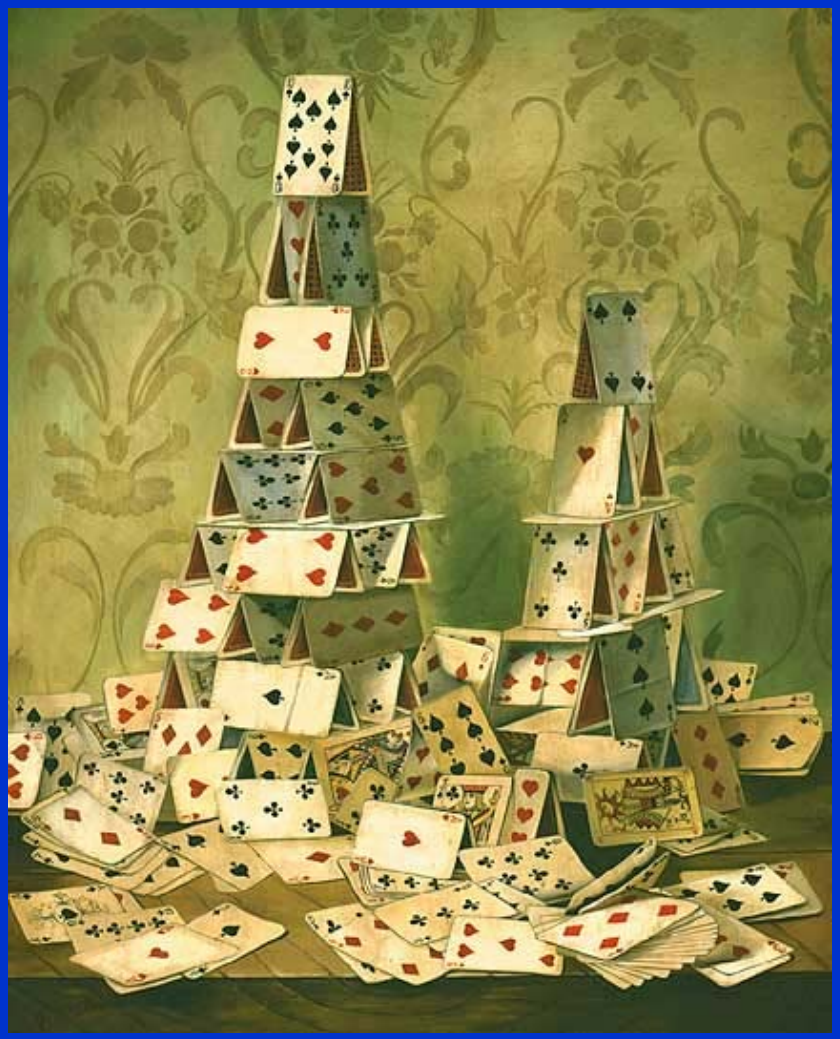
See Comment page 840

**Peninsula Clinical Trials Unit, Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth, UK** (Prof J Zajicek PhD, S Ball MSc, Prof D Wright PhD, J Vickery MSc, M Gomez Cano PhD, Prof J Hobart PhD); **Medical Research Council Clinical Trials Unit, London, UK** (Prof A Nunn MSc); **and Institute of Neurology, University College London, London, UK** (Prof D Miller FMedSci, D McManus MSc, S Mallik MSc)

Correspondence to:
Prof John Zajicek, Room N13, ITTC Building 1, Tamar Science Park, Plymouth PL6 8BX, UK
**john.zajicek@nhs.net**

# Importance...


Clinical Practice
Implications
Results
Statistics
Measurement



# Recommendations for advancing guidance

- Focus on conceptual clarity of variables for measurement
- Encourage application of clinically meaningful modern psychometric methods in COA development, evaluation, selection, and trial data analysis
- Within-trial instrument evaluation
- Consistency with other regulatory bodies