On the Measurement of Change in Medical Research

Ronir Raggio Luiz[1] and Renan Moritz V.R. Almeida[2,*]

[1] *Instituto de Estudos em Saúde Coletiva/Universidade Federal do Rio de Janeiro, Brazil*

[2] *Programa de Engenharia Biomédica, COPPE/Universidade Federal do Rio de Janeiro, Brazil*

Abstract: Measuring of change is essential in medical research. However, these measurements may have different goals and, traditionally, the ability to measure change has focused on sensitivity in a statistical sense, whereas little attention has been directed to the appropriate interpretation and analysis of change indicators. The present report examines some of the most important issues involved in measuring change with pre and post-test data when ordinal scales are used, and the conceptual problems pertaining to the use of these scales are also discussed. It can be said that there is still no agreement about the most adequate strategy for assessing health status change in a group of subjects, what caused the introduction of many indicators, most of which variations of the ES (Effect size: the mean of change scores divided by the standard deviation of the baseline scores) concept. The adequate interpretation of change scores in these cases demands a high degree of knowledge about what these changes mean to specific sub-groups of patients, as well as detailed information on their situation at baseline, such as score distributions. Researchers should strive for interpretations that take into account what "change" means for different patients.

Keywords: Measurement, measurement scales, ordinal variables, change indicators, effect size.

INTRODUCTION

Measuring change is essential for scientific research [1]. However, these measurements may have different goals (Box **1**); and, traditionally, the ability to measure change has focused on sensitivity in a statistical sense, whereas little attention has been directed to the appropriate interpretation and analysis of change indicators. In addition, many researchers do not realize that measurement problems, e.g. when ordinal variables are involved, are not completely settled in the literature [2-7].

The present report examines some of the most important issues involved in measuring change with pre and post-test data when ordinal scales are used. Some of the conceptual problems pertaining to the use of these scales are also discussed. More detailed reviews of the problems introduced by the use of ordinal scales may be found, for instance, in [4, 8-11].

MEASUREMENT OF CHANGE

The most basic strategy to measure change in an observational unit consists in determining two time points (pre or initial and post or final the intervention) and calculating the difference between $X_{initial}$ and X_{final}. That difference is known as change score, and most measures of change are calculated by indicators that have mean change for the total group in their numerators and a measure of variability in their denominators. However, Liang [12] differentiates between sensitivity to change, (the ability to measure any degree of change) and responsiveness (the ability to measure clinically important change), and Terwee *et al.* [13] identified 25 operational definitions for the latter, based on the ability to detect clinically important change or to detect real change (change that takes into account a gold standard). Concerning the former, the most common approaches were: the Effect size (ES), the Standardized Response Mean (SRM) and the Guyatt´s responsiveness statistic [14] (Table **1**).

ES is defined as the mean of change scores divided by the standard deviation of the baseline (initial) scores. The SRM and the ES are conceptually similar, so that the SRM is the mean change divided by standard deviation of the change scores. Finally, Guyatt's responsiveness statistic, also a variant of the ES, is the Minimal Clinically Important Difference (MCID) divided by an estimate of the within-individual variability for subjects who are stable, in the case of two measurements (e.g., before and after an intervention) [15]. When the correlation between baseline and follow-up scores is equal to 0.5, ES equals SRM; when it is higher, SRM is greater than ES; and when it is lower, ES is higher than SRM [16].

SOME CONCEPTUAL ISSUES IN THE MEASUREMENT OF CHANGE

All three measures described above are usually interpreted considering the same benchmarks (0.2 or less: small, 0.5: moderate, and 0.8 or greater: large) [17]. As an example, the indexes above are presented with the help of a hypothetical study in which ten

*Address corresponding to this author at the Programa de Engenharia Biomédica, COPPE – Universidade Federal do Rio de Janeiro, Caixa Postal 68510 Cidade Universitária Rio de Janeiro RJ, Brazil; Tel: +(55)21 25628583; Fax: +(55)2125628591; E-mail: renan@peb.ufrj.br

E-ISSN: 1929-6029/12

Author (s)	Objectives
Liang [11]	To distinguish any level of clinically important change
Husted *et al.* [16]	To evaluate repeated measures in a group before and after a treatment To evaluate the relationship between the variation relatively to a gold standard
Norman *et al.* [11]	To evaluate change taking the gold standard as the base To identify the change between treatment group and control group
Terwee *et al*. [12]	To detect change in general To detect clinically important change To detect real change in the concept being measured

Box 1: Possible Objectives of Measuring Change.

patients with painful cancer disease were measured at two times (Table **2**). The MCID was estimated considering the average change score among those patients rating some improvement minus the average change score among those patients rating no change [17]. The change value defined as clinically significant was 0.4. The results for each indicator are ES = 0.98; SRM = 1.14 and Guyatt's = 0.09, and, thus, ES and SRM represent a change considered as large, while Guyatt's indicates a small change.

However, Guyatt's is strongly influenced by the definition of MCID, which may not be straightforward. For instance, "clinical importance" can be defined as "usual" improvement rates for a type of patients, as targets for improvement or as expected physiological changes. Ostelo *et al.* [15] identified a range of values in MCID from 1 to 4.5 (absolute values) in a study for pain with a Numerical Rating Scale (scoring range 0–10). In the "Table **2**" example, if MCID were 4.5, Guyatt's would then take the value 1.06 (even higher than 0.8). Therefore, the use of the same benchmarks under these circumstances actually indicates a confusion and lack of rigour in the measurement literature, introducing a strong possibility of result misinterpretation.

Two other important issues that influence the interpretation of change scores are the regression toward the mean and the measurement problem produced by the use of ordinal scales [14]. Regarding the first, subjects with high values at baseline ($X_{Initial}$) may migrate to lower values at (X_{Final}), while, conversely, subjects with low values migrate to higher values. Then, the movement of subjects may be interpreted as resulting from the intervention, when in fact it is related to the high correlation between change and baseline values. This is of especial concern when subjects are selected at baseline according to their scores. For instance, if subjects with high cholesterol levels start a treatment, some of them will move towards lower values ("the mean") whatever the effect of the treatment. An option for dealing with this problem is the use of linear regression models to estimate the correlation between baseline and follow-up scores [18].

Concerning ordinal scales, their definition intrinsically implies that the distance between each class in the scale is not known, and the information thus produced is difficult to interpret and easy to misuse. The seminal work of Stevens [19] defined a typology of variables according to their measurement scales, classifying them as nominal, ordinal, interval, and ratio, an idea that stressed the relationship between measurement scales and adequate statistical methods. As a consequence of Stevens ideas, two groups of thought were created: the liberal (anti-Stevens) and the conservative (pro-Stevens). For the former, differences among the categories of the ordinal scale are the same, and mathematical operations are, therefore, possible. However, it is important to remember that the number of categories of the

Table 1: Most Important Indexes for Measurement of Change [13]

Measure	Defined as...
Effect size (ES)	[Mean ($X_{Initial}$ – X_{Final})] / $\sigma_{Baseline}$
Standardized response mean (SRM)	[Mean ($X_{Initial}$ – X_{Final})] / $\sigma_{Difference}$
Guyatt's responsiveness statistic (GRS)	MCID / √ 2 * MSE_X

In this table, X represents the observed values, $X_{initial}$ the baseline values and X_{final} the values after treatment, $\sigma_{Baseline}$ is the standard deviation of baseline scores, $\sigma_{Difference}$ the standard deviation of difference scores, MICD the minimally clinically important change and MSE_X the Mean Square Error of X obtained from an ANOVA.

Table 2: A Hypothetical Example of Ten Patients Assessed on a Pain Scale (Numerical Rating Scale) in Two Moments of Time

Patient	Initial Assessment	Follow-up Assessment	Change: Initial - Follow-up
1	0	0	0
2	8	5	3
3	10	1	9
4	3	2	1
5	3	8	-5
6	5	1	4
7	8	0	8
8	0	2	-2
9	2	0	2
10	6	4	2

Notice that the scoring of patient 10 changed from 6 to 4, indicating progress, and the scoring of patient 6 changed from 5 to 1, indicating greater progress, but the “5 to 1” change does not necessarily represent twice as much progress.

analyzed variables should also be considered. It is reasonable to accept that, with a small number of labels, one should rather stay at the conservative (pro-Stevens) side, but, as the number of labels increases, the liberal (anti-Stevens) side becomes more and more appropriate."

As an example of the problems thus introduced, consider again the example on Table **2**. Movement from 6 to 4 (patient 10) represent progress in pain management, and from 5 to 1 (patient 6) greater progress, but not necessarily twice as much progress, since a gain of 2 has a different meaning for patients with poor versus good baseline status. Therefore, the following questions may arise: what does a variation of 2 points in a pain evaluation scale mean? Do the results obtained with individuals 2 points above the minimum have the same interpretation, even if one knows that they began the study with different scores? Is a 5-point variation of an individual significantly different from the 7-point variation of another? How is it possible to establish a hypothesis of difference equal to zero, among consecutive measures, if the values are not numbers but numeric labels? It follows that in contrast to interval and ratio scales, ordinal numbers are at best symbols of “greater than” and “less than” quantities, and have even been called “non-numbers” [8].

CONCLUSION

Unbeknown to many researchers, there is still no agreement about the most adequate strategy for assessing health status change in a group of subjects. This situation caused the introduction of many indicators, both old and new [20], most of which are variations of the ES concept. Furthermore, change results that can be deemed “large” or “small” can be obtained simply by a convenient choice of indicator, and, at present, there is no consensus on the best method for estimating an index such as the MCID.

Additionally, ordinal scales offer a fast and inexpensive way to characterize complex phenomena, but they can be misleading. Adequate interpretation of change scores in the case when ordinal scales are used demands a high degree of knowledge about what these changes mean to specific sub-groups of patients, as well as detailed information on their situation at baseline, such as score distributions. One should keep in mind that researchers should strive for interpretations that take into account what “change” means for different patients.

ACKNOWLEDGEMENTS

A substantial contribution to the present work was made by our friend and colleague Mario Luiz Pinto Ferreira, who unexpectedly died prior to the study conclusion.

APPENDIX: SYMBOLS AND ABBREVIATIONS USED

ES	= Effect size; the mean of change scores divided by the standard deviation of the baseline scores

SRM = Standardized Response Mean; the mean change divided by standard deviation of the change scores

MCID = Minimal Clinically Important Difference

X = measurement values

$X_{initial}$ = baseline measurement values

X_{final} = values after treatment

$\sigma_{Baseline}$ = standard deviation of baseline scores

$\sigma_{Difference}$ = standard deviation of difference scores

MSE_X = Mean Square Error of X obtained from an Analysis of Variance (ANOVA)

REFERENCES

[1] Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997; 50(3): 239-46. http://dx.doi.org/10.1016/S0895-4356(96)00373-3

[2] Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. J Clin Epidemiol 1996; 49(7): 711-17. http://dx.doi.org/10.1016/0895-4356(96)00016-9

[3] Svensson E. Comparison of the quality of assessments using continuous and discrete ordinal rating scales. Biometr J 2000; 42(4): 417-34. http://dx.doi.org/10.1002/1521-4036(200008)42:4<417::AID-BIMJ417>3.0.CO;2-Z

[4] Kampen J, Swyngedouw M. The ordinal controversy revisited. Quality Quantity 2000; 34: 87-102. http://dx.doi.org/10.1023/A:1004785723554

[5] Cohen ME. Analysis of ordinal dental data: evaluation of conflicting recommendations. J Dent Res 2001; 80(1): 309-13. http://dx.doi.org/10.1177/00220345010800010301

[6] Michell J. The psychometrician´s fallacy: Too clever by half? Br J Math Statist Psychol 2009; 62: 41-55. http://dx.doi.org/10.1348/000711007X243582

[7] Kemp S, Grace RC. When can information from ordinal scale variables be integrated? Psychological Methods, 2010; 15(4): 398-12. http://dx.doi.org/10.1037/a0021462

[8] Merbitz C, Morris J, Grip JC. Ordinal Scales and foundations of misinference. Arch Phys Med Rehabil 1989; 70: 308-12.

[9] Knapp TR. Treating ordinal scales as interval scales: an attempt to resolve the controversy. Nursing Res 1990; 39: 121-23.

[10] Velleman PF, Wilkinson L. Nominal, ordinal, interval, and ratio typologies are misleading. Am Statistic 1993; 47(1): 65-72.

[11] Norman GR, Sridar FG, Guyatt GH, Walter SD. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. Med Care 2001; 39: 1039-47. http://dx.doi.org/10.1097/00005650-200110000-00002

[12] Liang MH. Longitudinal construct validity. Establishment of clinical meaning in patient evaluative instruments. Med Care 2000; 38(9)(Suppl II): II-84 - II-90.

[13] Terwee CB, Dekker FW, Wiersinga WM, Prummel FM, Bossuyt PMM. On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. Quality Life Res 2003; 12: 349-62. http://dx.doi.org/10.1023/A:1023499322593

[14] Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Reprinted 2005. New York: Oxford University Press 2005; pp. 196-212.

[15] Ostelo RWJG, Deyo RA, Stratford P, Waddell G, Croft P, von Korff M, *et al*. Interpreting change scores for pain and functional status in low back pain. Spine 2008; 33(1): 90-94. http://dx.doi.org/10.1097/BRS.0b013e31815e3a10

[16] Fortin PR, Stucki G, Katz JN. Measuring relevant changes: An emerging challenge in rheumatologic clinical trials. Arthritis Rheum. 1995; 38: 1027-30. http://dx.doi.org/10.1002/art.1780380802

[17] Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000; 53: 459-68. http://dx.doi.org/10.1016/S0895-4356(99)00206-1

[18] Cronbach LJ, Furby L. How should we measure change or should we? Psychol Bull 1970; 74: 68-80. http://dx.doi.org/10.1037/h0029382

[19] Stevens SS. On the theory of scales of measurement. Science 1946; 103: 677-80. http://dx.doi.org/10.1126/science.103.2684.677

[20] Ferreira MLP, Almeida RMVR, Luiz RR. A new indicator for the measurement of change with ordinal scores. Quality Life Res, in press.

Received on 20-10-2012 Accepted on 22-11-2012 Published on 21-12-2012

http://dx.doi.org/10.6000/1929-6029.2012.01.02.07