2016-06-13

Accounting for test reliability in student progression: the reliable change index

Zahra, Daniel

http://hdl.handle.net/10026.1/5235

10.1111/medu.13059
Medical Education
Wiley

This is a final author's draft of a paper submitted for publication in Medical Education 2016 DOI: http://dx.doi.org/10.1111/medu.13059

Running Head: Student Progression and Reliable Change

Accounting for test reliability in student progression: The Reliable Change Index

Abstract

Developed by Jacobson and Truax (1), the reliable change index (RCI) provides a measure of whether the change in an individual's score over time is within or beyond what might be accounted for by measurement variability. In combination with measures of whether an individual's final score is closer to one population or another, this provides useful individual-level information which can be used to supplement traditional analyses. This article aims to highlight its potential for use within medical education, and in particular as a novel means of monitoring progress at the student-level across successive test occasions or academic years. We provide an example of how it can be applied informatively to assessment evaluation and discuss its wider usage. This approach can be used to identify and support failing students as well as to determine best teaching and learning practices by identifying high-performing students. Furthermore, the individual-level nature of the RCI makes it well suited for educational research with small cohorts, as well as tracking individual profiles within a larger cohort or addressing questions about individual performance that may be unanswerable at the group-level.

Introduction

The Reliable Change Index (RCI), developed by Jacobson and Truax (1) over twenty years ago, provides a way of capturing not only the *statistical* but also *clinical* significance of a change over time – and most importantly, after taking into account the reliability of the measures used to capture the change (2). Although the RCI was originally developed for use in the medical field, it has great value to other areas of research as well. Zahra and Hedge (3) for example discuss the applicability of the RCI to academic psychology as a measure of individual progression over time. The authors, however, highlight the fact that as in many disciplines group-level analyses such as ANOVAs and t-tests are favoured over individual level ones. This article aims to highlight the RCI's potential for use within medical education as a novel means of monitoring progress at the student-level across successive test occasions or academic years.

Measuring change in individuals has been shown to be notoriously difficult in the area of educational assessment, with authors highlighting a range of issues (the following works are recommended for those seeking further discussion: 4, 5-11). There are many obstacles to determining the extent that individuals learn to greater or less extents than others. Measurement is never perfect, and educators face the challenge of evaluating meaningful change in the presence of noise. The technique we discuss does not remove these concerns, but by characterising the quality of student's assessment scores, it allows assessors to make the best use of the information that is available to them.

The RCI in Medical Assessment

In medical education, particularly in assessments such as progress tests, it is important to track student scores over time. Of most interest is perhaps whether students are improving year on year or test on test as they progress through their degrees. Assuming you have a cohort who have completed two tests measuring related content, for example exams a medical knowledge test at the start and end of an academic term (Test 1 and Test 2 for purposes of illustration), you might run a t-test on the means of each exam in order to evaluate progression and report something like "test scores in Test 2 are significantly higher than they were in Test 1, $t(54)=-5.38$ $p<.001$". You might even say that "the improvement was large in terms Cohen's (12) effect size, d=0.86".

But that is statistical significance as based on the mean performance of each group. Such an extreme difference is unlikely to be due to chance changes in student knowledge (13), but it tells us very little about how meaningful that change is, or how each individual student has progressed. It doesn't allow us to make statements that are meaningful in terms of how one particular student is performing in relation to the rest of the cohort at the time of Test 1 or Test 2 – is their performance, even in their second test, closer to the cohorts performance on the first test, or are they 'keeping pace' with their peers? Being able to address questions like these has a range of applications in medical education, from identifying struggling students for remediation to identifying those outperforming their current or even senior year groups.

In clinical work, this is the idea of *clinical* significance (1). In considering student progress, not only is there an interest in overall group - or individual - change from a statistical point of view, but what is critical is whether the individual is closer to one group or another, be that a control group in a clinical trial, or a year group in a knowledge test. A change is *clinically* significant if the individual or group has moved from being more like one population to being more like another, where 'more like' can be defined as a given score being probabilistically more likely to belong to an individual in one group rather than the other. In our example, a student will have shown 'clinically' significant (in this case, perhaps better thought of as 'educationally' significant) change if they progress from being closer to the Test 1 score distribution to being closer to the Test 2 score distribution.

Yet another factor to consider in such settings is the reliability of the change. Can the change be accounted for by variability in the measures being used, the reliability of the exam? Unfortunately this is where the standard tests start to become of less use to students and educators, but where these considerations are explicitly included in the reliable change index.

Calculating Reliable Change Indices

The focus of the RCI is on individual change over time, not changes in overall group performance. In the context of medical education this is change at the student level; whether an individual student is improving , whether that improvement is reliable, and finally, whether that change puts the student closer to the performance of one year-group or another. In practical terms, although popular statistics packages such as SPSS and STATA don't typically provide reliable change measures, they are relatively easy to compute. Equation 1 shows the calculation of RCI scores based on a combination of the equations published by Jacobson and Truax (1).

Equation 1: $$RCI = \frac{x_2 - x_1}{\sqrt{2\left(s\sqrt{1-r_{xx}}\right)^2}}$$

Where x_1 and x_2 are an individual student's scores for Test 1 and Test 2, s is the standard deviation at the first time-point, and r_{xx} is the test-retest reliability (in our example; though see 'Estimating Reliability of Tests' for further discussion of reliability estimates for the RCI).

In other words, the top of the formula reflects the change in an individual's performance, and the bottom of the formula captures the degree of noise in the measure. As a result, as the reliability goes down, the value of the lower half increases (i.e. it's harder to detect change). Therefore, reliable measurement is still a principle concern. This highlights a key question in the use of the RCI which will be discussed in detail below, namely, how to calculate and incorporate an estimate of reliability across two exams.

The direction of change, its size, and its reliability are captured by the RCI. An RCI score of 1.00 is a change half the size of an RCI of 2.00, and RCI scores with a magnitude of 1.96 or greater can be considered statistically significant at the $p<.05$ level (1). RCI scores greater in magnitude than 1.96 represent a change over and above what might be accounted for by the variability of the measure. RCI scores with a magnitude less than 1.96 may be 'real' changes, but they may also be accounted for by measurement variability. This margin of reliability in relation to examination scores provides an area within which changes might be due to measurement variability, and potentially not reflect true improvement (or deterioration) in performance. This is explained below.

With respect to how meaningful any individual's change is, Jacobson and Truax (1) provide a detailed discussion of methods by which a 'clinical' significance cut-off can be determined, but it essentially provides a threshold indicating which distribution of scores the student is closest to, or more representative of. In the case of yearly exams, the two score distributions can be treated as curves. If a normal distribution is assumed, with Test 1 scores having M=47.00 and SD=16.27, and Test 2 scores having M=60.91 and SD=16.27, the simplest criteria for 'clinical' significance is the mid-point of the two means. If equal variances can be assumed, this is calculated as shown in Equation 2.

Equation 2: $$midpoint = \frac{(M_1+M_2)}{2}$$

If equal variance cannot be assumed, the criteria for clinical significance can be calculated as in Equation 3 where M_1 and M_2 are the means of the two distributions, and s_1 and s_2 are the standard deviations (for a more detailed discussion, see reference 1).

Equation 3: $$midpoint = \frac{s_1M_1+s_2M_3}{s_1+s_2}$$

The information provided by the RCI and clinical cut-off point can be combined and presented as in Figure 1 for easy reference by staff and students. When plotting the scores from Test 1 against the scores from Test 2, the heavy diagonal line shows points of no-change. Anyone above this has improved their score, anyone below it has seen a decrease in their score. Change that could be accounted for by variation in the test is bounded by the two thin diagonal lines, whereas scores outside of this diagonal swathe have RCI>|1.96| and thus show reliable change. The mid-point between Test 1 and Test 2 score distributions (using the midpoint between the means) is indicated by a dashed horizontal line.

[Figure 1]

In interpreting this representation, Molly has scored higher in Test 2 than in Test 1 (above the $y=x$ diagonal). Her improvement is reliable; above what might be expected due to measurement variability (above the upper diagonal), and puts her closer to the Test 2 distribution than the Test 1 distribution (above the dashed horizontal line). Despite showing reliable improvement, James' score is closer to the distribution of first-test scores than second-test scores. Ahmed and Tom both show improvement between tests one and two, but improvements which may be accounted for by the variability of the measurement. Furthermore, Ahmed's Test 2 score places him closer to the second-test distribution, whereas Tom's remains closer to the first-test scores, suggesting a lack of genuine improvement from the start of the year. Jago, Charlotte, and Sarah are potentially doing less well. Their scores have all decreased between Test 1 and Test 2. Despite this, Jago is just over the dashed line, and still remains closer to the Test 2 distribution; and both Jago and Charlotte's progress is still within the bounds of measurement variability. Sarah, however, has performed more poorly in Test 2 than in Test 1, has shown a decrease outside the bounds of measurement variability, and is ultimately closer to the Test 1 distribution than the Test 2 distribution.

Estimating the Reliability of Tests

As mentioned above, the RCI takes into account the reliability of the measure being used. Initially the RCI was developed to incorporate the test-retest reliability of a measure when that measure was used to evaluate change over time in a particular construct. The most straightforward application of this approach would be instances in which the same test questions are administered at multiple time points. Where this is not possible, assume that the two administrations (i.e. Test 1 and Test 2) reasonably represent parallel forms of the same test. This is most applicable when the different test-occasions reflect a common construct, but is not a trivial assumption, and should be empirically validated where possible.

Even in progress test situations, that students will be sitting the same test on multiple occasions is unlikely. However, as the tests are designed to measure the same construct, such as applied medical knowledge for example, the test-retest reliability can be incorporated as the correlation between the two test occasions (Test 1 and Test 2 in our example). It is important to consider what is being assessed by the tests at each time-point when considering use of the RCI. In progress tests, or knowledge tests, where the content across all tests is drawn from a pool of 'all knowledge covered by the curriculum', test-retest reliability can be used. Where the tests measure different constructs, or only subsets of items measure the same construct it may be more appropriate to create subsets of these items for analysis of reliable change. Educationally, development in knowledge within a domain or topic area is usually of most interest, and it is these instances of retesting common constructs to which the RCI can add an additional dimension of understanding. As discussed above, for example, identifying particularly excelling students, or identifying students who are struggling to develop their knowledge within a particular domain.

Related to the incorporation of the reliability is the calculation of the standard error of measurement in the RCI formula presented above (Equation 1), derived from the work of Jacobson and Truax (1). In Equation 1, the element incorporating this is:

Equation 4: $S_E = s\sqrt{1 - r_{xx}}$

However, Maassen (14) highlights methods of calculating this which may be considered less reliant on distributional assumptions. As traditional assessment analyses typically rely on assumptions such as normality, we have focussed on and presented examples using the Jacobson and Truax (1) formulae, but would recommend Maassen's (14) work to the interested reader or those who routinely work with skewed data. Similarly, given the RCI's focus on change over time, regression to the mean may be an issue. In such cases, we suggest the RC_{ID} formula proposed by Hageman and Arrindell (15). The calculation of these is more complex, but those wishing to explore the robustness of the RCI in relation to multiple test occasions are likely to find their discussions valuable.

Usefulness as a research and educational tool

Most research is conducted at the group level and is focussed on testing hypotheses which can be generalised to a wider population, but where the interest is on individuals or smaller subgroups the RCI is a useful tool for both research and education. It can be used to evaluate interventions designed to improve the learning and experiences of subsets within cohorts, for which analysis at an individual level is perhaps more appropriate.

Although we would not argue that the RCI is by any means a replacement for group-level analysis (e.g. see Alternative Approaches and Limitations section), the RCI also provides a useful means of giving individual level feedback to students, especially as it overcomes the typically prohibitive demands on resources for producing individualised feedback. This may be of particular value in quantitative examinations such as assessments over multiple test occasions, or providing feedback on progress from year to year. This is particularly true as the approach allows accessible visual feedback and there is body of work which suggests that individual feedback in any form is of much more use to the student than group-level feedback. Furthermore, this focus on individual change and ability to track and quantify individual progress are lacking in more common statistical approaches. The RCI therefore allows educators to identify struggling students who may otherwise be overlooked if they show improvement, which is not reliable, and hence tailor support tools for these students.

When samples or cohorts are very small and group-level analyses are not feasible, the RCI provides an alternative approach. It lends itself to research or feedback in situations involving small subsets of students (i.e., ethnic sub-groups; individuals with learning disabilities), but can still be applied to larger cohorts. As suggested above, this is particularly useful when the focus is on specific populations or subgroups which may have smaller memberships or be difficult to recruit from. Because of the focus on the individual rather than the group as a whole, RCI scores also allow classification of people into those whose performance has been reliably altered by an intervention and those whose scores have not, as well as the identification of performance profiles which are unusual.

However, given its dependence on reliability, which may be influenced by a number of factors, we would not suggest it be used for high-stakes decisions; its strengths lie in supplementing more routine analyses.

In Practice

Although the RCI provides a useful tool for tracking individual change, in practice, there are very few instances where this is the only topic of interest. However, in most research designs the RCI can be used alongside traditional tests in order to add a further dimension to the findings and our understanding of the data. With respect to progress testing for example, changes in performance between two time-points might be visualised using scatter plots. This could be augmented by including boundaries for reliable change. In addition, the performance of students who reliably improve or deteriorate could complement cohort item analysis data in order to provide a clearer indication of which groups of students perform well or less well on particular questions, or to inform discussions during item review.

The use of the RCI to augment more traditional approaches to assessment and evaluation, alongside discussion of its uses in the realm of remediation, highlights the need to keep in mind the practical application of the RCI, and traditional statistics more generally, when making decisions related to student assessment. In particular, regression to the mean may explain particularly sudden increases or decreases in student scores. Although this can be controlled for to some extent when processing consecutive examinations (16), consideration of the reliability of changes over longer periods might help to provide a more accurate and robust picture of a student's progression. Furthermore, factors such as low scores which need to be discounted due to extenuating circumstances need to be thoroughly checked as outlying data could skew the distribution and in turn skew RCI calculations.

An additional point to consider in using the RCI to track change over multiple exams is practice effects. Although 'practice' is a necessary characteristic in education, to what extent similarities between exams might account for performance changes is worth considering before drawing conclusions from the results. Although identical exams are unlikely to be administered on multiple occasions to the same students, the correction for practice effects discussed by Chelune, Maugle, Lüders, Sedlak and Awad (17) and Temkin, Heaton, Grant, and Dikmen (18) might be considered if the investigator were interested in trying to minimise the influence of potential practice effects or changes in exam strategy though they may often be confounded.

Implementation in Excel, R, and STATA

As mentioned above, RCI procedures are not included as standard in popular statistical software packages. However, applying the equations provided here in packages such as Excel, R, and SATA is relatively straightforward.

For presentation to students and staff, the information provided by the RCI and cut-off points is perhaps best presented graphically, as in Figure 1. The package ggplot2 (19) was used in R to create the image for the current paper, but similar results can be achieved with other packages, and with other software. Resources for implementing the RCI using R, STATA, and Excel can be found at http://tinyurl.com/pppxwr3, and the authors are more than happy to discuss these resources (contact details above).

Alternative Approaches and Limitations

As we have noted in this article, the RCI is not a substitute for reliable measurement, which is a particular concern in the measurement of change. What it does do is provide a relatively simple method for accounting for the quality of measures, and uses this information as a tool to help students and help educators improve assessment and practice. The RCI also takes the individual as the unit of interest, supplementing analysis questions which typically assess performance differences at a group level.

The RCI is grounded in the assumptions of classical test theory (20) , which is likely the context within which most readers understand issues of measurement reliability and measurement error. Numerous other techniques for assessing change also exist (21). One prominent alternative approach is that of Item Response Theory (IRT; 22); which comprises a framework and set of techniques for dissociating both individuals and items (e.g. test questions) with respect to one or more underlying “latent” dimensions. These approaches show a great deal of merit, though are not without limitations and caveats, such as the accessibility of the techniques and statistical software required to implement them. We would direct the interested reader to recent reviews on these approaches (e.g. 23).

In spite of the benefits of considering the RCI there are, as with all analyses, some caveats to consider. Many of these have already been discussed, but we reiterate the key ones here. Firstly, estimates of reliability need to be carefully considered. The RCI was developed to incorporate test-retest reliability when evaluating change over time, but given the nature of educational assessments, correlations between the two test occasions may be the best available indicator of reliability. Secondly, as with typical analyses, the RCI may also be influenced by other forms of measurement error, regression to the mean, and practice effects. There are various adjustments that have been proposed for accounting for these (e.g. reference 15 introduced above), but it is up to the researcher to determine if the required data can be obtained, and if the potential reduction in biases outweighs the added complexity. Finally, careful consideration of all assessment analyses should be used collectively. Information from the individual-level RCI data can inform interpretation of group-level analyses and vice-versa to reach defensible and robust conclusions about student progression.

Summary

In summary, the reliable change index has a number of potential uses in medical education, particularly when the individual is the focus of consideration. It is particularly suited to analysis of change over time in individuals, or small subgroups which cannot be captured using the more common analysis techniques. In these instances it provides a simple way of accounting for the reliability of the measure and can indicate where change is over and above the variability of the measurement tool. Although careful thought is needed with respect to derivation of parameters needed in its calculation and whether subsequent adjustments for other forms of measurement error are needed, it provides a means of gaining more from routinely produced assessment data. This can then be used to improve and inform decisions relating to student growth, assessment design, and student feedback.

References

1. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology. 1991;59(1):12-9.

2. Jacobson NS, Roberts LJ, Berns SB, McGlinchey JB. Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. Journal of Consulting and Clinical Psychology. 1999;67(3):300-7.

3. Zahra D, Hedge C. The reliable change index: Why isn't it more popular in academic psychology? Psychology Postgraduate Affairs Group Quarterly. 2010;76:14-9.

4. Cronbach LJ, Furby L. How we should measure change - or should we? Psychological Bulletin. 1970;74(1):68-80.

5. Lord FM. The measurement of growth. Educational and Psychological Measurement. 1956;16(1):421-37.

6. Edwards JR. Ten difference score myths. Organizational Research Methods. 2001;4(3):265-87.

7. Rogosa D. Myths about longitudinal research. In: Schaie KW, Campbell TR, Meredith W, Rawlings SC, editors. Methodological research in ageing research. New York: Springer; 1988. p. 171-210.

8. Rogers PJ. Myths and methods: "Myths about longitudinal research" plus supplemental questions. . In: Gottman JM, editor. The analysis of change. Hillsdale, New Jersey.: Lawrence Erlbaum Associates.; 1995. p. 3-65.

9. Willet JB. Questions and answers in the measurement of change. Review of Educational Research. 1988;15(1):345-422.

10. Zimmerman DW, Williams RH. Reliability of gain scores under realistic assumptions about properties of pretest and posttest scores. British Journal of Mathematical and Statistical Psychology. 1998;51(2):343-51.

11. Zumbo BD. The simple difference score as an inherently poor measure of change: Some reality, much mythology. In: Thompson B, editor. Advances in Social Science Methodoloy Greenwich, Connecticut: JAI Press; 1999. p. 269-304.

12. Cohen J. Statistical power analysis for the behavioural sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

13. Field A. A bluffer's guide to effect sizes. Postgraduate Affairs Group Quarterly. 2006;58(March):9-23.

14. Maassen GH. The standard error in the Jacobson and Truax reliable change index: the classical approach to the assessment of reliable change. Journal of The International Neuropsychological Society. 2004;10:888-93.

15. Hageman WJJ, Arrindell WA. A further refinement of the reliable change (RC) index by improving the pre-post difference score: introducing RC_{ID}. Behavior Research and Therapy. 1993;31(7):693-700.

16. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. International Journal of Epidemiology. 2005;34:215-20.

17. Chelune GJ, Naugle RI, Lüders H, Sedlak J, Awad IA. Individual change after epilepsy surgery: practice effects and base-rate information. Neuropsychology. 1993;7(1):41-52.

18. Temkin NR, Heaton RK, Grant I, Dikmen S. Detecting significant change in neuropsychological test performance: a comparison of four models. Journal of The International Neuropsychological Society. 1999;5:357-69.

19. Wickham H. Ggplot2: Elegant graphics for data analysis. New York: Springer; 2009.

20. Novick MR. The axioms and principal results of classical test theory. Journal of Mathematical Psychology. 1966;3(1):1-18.

21. Bauer S, Lambert MJ, Neilsen SL. Clinical significance methods: a comparison of statistical techniques. Journal of Personality Assessment. 2004;82(1):60-70.

22. Lord FM. Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Lawrence Erlbaum; 1980.

23. Thomas ML. The value of Item Response Theory in Clinical assessment: A Review. Assessment. 2011;18(3):291-307.

Acknowledgements

The current work adapts the ideas presented by [removed for anonymous review] for application within a medical education environment. We are indebted to the many individuals who have contacted us over the years to discuss the reliable change index in relation to a wide range of research areas, and are grateful to them for their suggestions and ideas regarding its relevance to medical education and related fields.

Declarations of Interest

The authors report no declarations of interest.

Figures

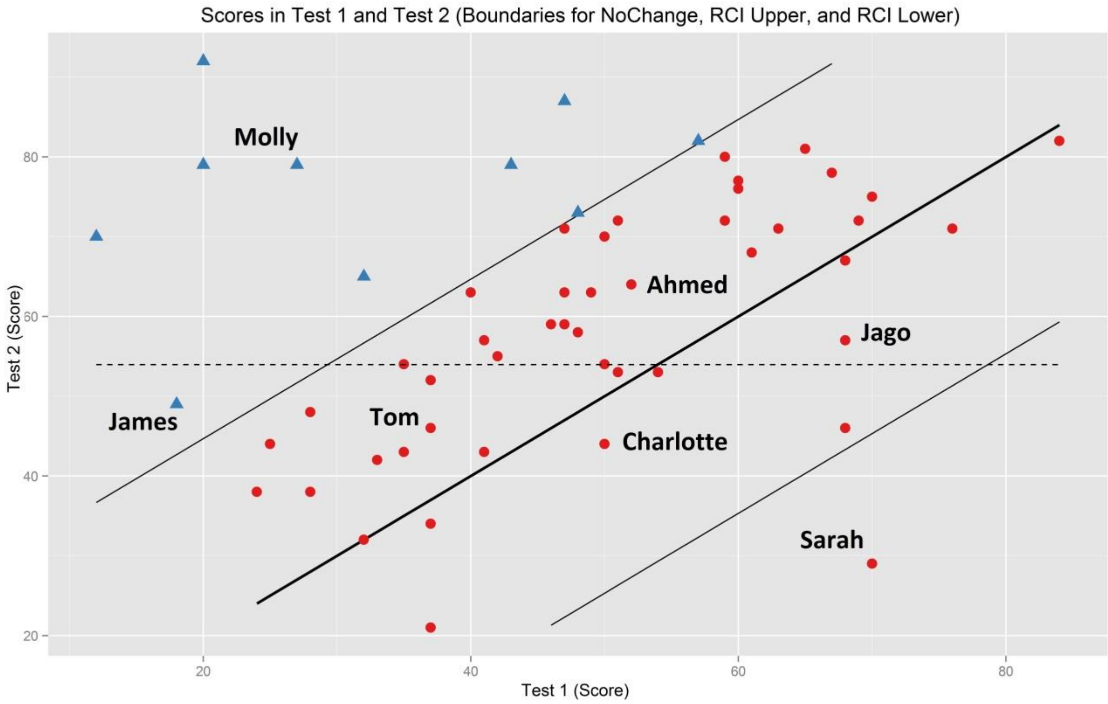

Figure 1: Scatter-plot showing Test 1 and Test 2 scores with a line of no-change (solid diagonal at y=x) upper and lower bounds for reliable change (narrow diagonals) and the mid-point of the means for each test (dashed horizontal)