

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/102763/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Zahra, Daniel, Hedge, Craig , Pesola, Francesca and Burr, Steven 2016. Accounting for test reliability in student progression: the reliable change index. *Medical Education* 50 (7) , pp. 738-745. 10.1111/medu.13059

Publishers page: <http://dx.doi.org/10.1111/medu.13059>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



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

1
2
3 year on year or test on test as they progress through their degrees. Assuming you have a
4
5 cohort who have completed two tests measuring related content, for example exams a
6
7 medical knowledge test at the start and end of an academic term (Test 1 and Test 2 for
8
9 purposes of illustration), you might run a t-test on the means of each exam in order to
10
11 evaluate progression and report something like “test scores in Test 2 are significantly higher
12
13 than they were in Test 1, $t(54)=-5.38$ $p<.001$ ”. You might even say that “the improvement
14
15 was large in terms Cohen’s (12) effect size, $d=0.86$ ”.

16
17
18
19 But that is statistical significance as based on the mean performance of each group.
20
21 Such an extreme difference is unlikely to be due to chance changes in student knowledge
22
23 (13), but it tells us very little about how meaningful that change is, or how each individual
24
25 student has progressed. It doesn’t allow us to make statements that are meaningful in terms
26
27 of how one particular student is performing in relation to the rest of the cohort at the time
28
29 of Test 1 or Test 2 – is their performance, even in their second test, closer to the cohorts
30
31 performance on the first test, or are they ‘keeping pace’ with their peers? Being able to
32
33 address questions like these has a range of applications in medical education, from
34
35 identifying struggling students for remediation to identifying those outperforming their
36
37 current or even senior year groups.
38
39
40
41
42

43
44 In clinical work, this is the idea of *clinical* significance (1). In considering student
45
46 progress, not only is there an interest in overall group - or individual - change from a
47
48 statistical point of view, but what is critical is whether the individual is closer to one group
49
50 or another, be that a control group in a clinical trial, or a year group in a knowledge test. A
51
52 change is *clinically* significant if the individual or group has moved from being more like one
53
54 population to being more like another, where ‘more like’ can be defined as a given score
55
56 being probabilistically more likely to belong to an individual in one group rather than the
57
58
59
60

1
2
3 other. In our example, a student will have shown 'clinically' significant (in this case, perhaps
4
5 better thought of as 'educationally' significant) change if they progress from being closer to
6
7 the Test 1 score distribution to being closer to the Test 2 score distribution.
8
9

10 Yet another factor to consider in such settings is the reliability of the change. Can the
11
12 change be accounted for by variability in the measures being used, the reliability of the
13
14 exam? Unfortunately this is where the standard tests start to become of less use to students
15
16 and educators, but where these considerations are explicitly included in the reliable change
17
18 index.
19
20

21 22 23 **Calculating Reliable Change Indices**

24
25 The focus of the RCI is on individual change over time, not changes in overall group
26
27 performance. In the context of medical education this is change at the student level;
28
29 whether an individual student is improving, whether that improvement is reliable, and
30
31 finally, whether that change puts the student closer to the performance of one year-group
32
33 or another. In practical terms, although popular statistics packages such as SPSS and STATA
34
35 don't typically provide reliable change measures, they are relatively easy to compute.
36
37 Equation 1 shows the calculation of RCI scores based on a combination of the equations
38
39 published by Jacobson and Truax (1).
40
41
42
43
44
45

$$46 \text{ Equation 1: } RCI = \frac{x_2 - x_1}{\sqrt{2(s\sqrt{1-r_{xx}})^2}}$$

47
48
49
50
51 Where x_1 and x_2 are an individual student's scores for Test 1 and Test 2, s is the standard
52
53 deviation at the first time-point, and r_{xx} is the test-retest reliability (in our example; though
54
55 see 'Estimating Reliability of Tests' for further discussion of reliability estimates for the RCI).
56
57
58
59
60

1
2
3 In other words, the top of the formula reflects the change in an individual's performance,
4
5 and the bottom of the formula captures the degree of noise in the measure. As a result, as
6
7 the reliability goes down, the value of the lower half increases (i.e. it's harder to detect
8
9 change). Therefore, reliable measurement is still a principle concern. This highlights a key
10
11 question in the use of the RCI which will be discussed in detail below, namely, how to
12
13 calculate and incorporate an estimate of reliability across two exams.
14
15

16
17 The direction of change, its size, and its reliability are captured by the RCI. An RCI
18
19 score of 1.00 is a change half the size of an RCI of 2.00, and RCI scores with a magnitude of
20
21 1.96 or greater can be considered statistically significant at the $p < .05$ level (1). RCI scores
22
23 greater in magnitude than 1.96 represent a change over and above what might be
24
25 accounted for by the variability of the measure. RCI scores with a magnitude less than 1.96
26
27 may be 'real' changes, but they may also be accounted for by measurement variability. This
28
29 margin of reliability in relation to examination scores provides an area within which changes
30
31 might be due to measurement variability, and potentially not reflect true improvement (or
32
33 deterioration) in performance. This is explained below.
34
35
36
37

38
39 With respect to how meaningful any individual's change is, Jacobson and Truax (1)
40
41 provide a detailed discussion of methods by which a 'clinical' significance cut-off can be
42
43 determined, but it essentially provides a threshold indicating which distribution of scores
44
45 the student is closest to, or more representative of. In the case of yearly exams, the two
46
47 score distributions can be treated as curves. If a normal distribution is assumed, with Test 1
48
49 scores having $M=47.00$ and $SD=16.27$, and Test 2 scores having $M=60.91$ and $SD=16.27$, the
50
51 simplest criteria for 'clinical' significance is the mid-point of the two means. If equal
52
53 variances can be assumed, this is calculated as shown in Equation 2.
54
55
56
57
58
59
60

1
2
3
4
5
6
7

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

8
9
10
11
12
13
14
15

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).

16
17
18
19

$$\text{Equation 3: } \textit{midpoint} = \frac{s_1 M_1 + s_2 M_2}{s_1 + s_2}$$

20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

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.

42
43
44
45

[Figure 1]

46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3 second-test scores. Ahmed and Tom both show improvement between tests one and two,
4
5 but improvements which may be accounted for by the variability of the measurement.
6
7 Furthermore, Ahmed's Test 2 score places him closer to the second-test distribution,
8
9 whereas Tom's remains closer to the first-test scores, suggesting a lack of genuine
10
11 improvement from the start of the year. Jago, Charlotte, and Sarah are potentially doing less
12
13 well. Their scores have all decreased between Test 1 and Test 2. Despite this, Jago is just
14
15 over the dashed line, and still remains closer to the Test 2 distribution; and both Jago and
16
17 Charlotte's progress is still within the bounds of measurement variability. Sarah, however,
18
19 has performed more poorly in Test 2 than in Test 1, has shown a decrease outside the
20
21 bounds of measurement variability, and is ultimately closer to the Test 1 distribution than
22
23 the Test 2 distribution.
24
25
26
27
28
29
30

31 **Estimating the Reliability of Tests**

32
33 As mentioned above, the RCI takes into account the reliability of the measure being used.
34
35 Initially the RCI was developed to incorporate the test-retest reliability of a measure when
36
37 that measure was used to evaluate change over time in a particular construct. The most
38
39 straightforward application of this approach would be instances in which the same test
40
41 questions are administered at multiple time points. Where this is not possible, assume that
42
43 the two administrations (i.e. Test 1 and Test 2) reasonably represent parallel forms of the
44
45 same test. This is most applicable when the different test-occasions reflect a common
46
47 construct, but is not a trivial assumption, and should be empirically validated where
48
49 possible.
50
51
52
53

54
55 Even in progress test situations, that students will be sitting the same test on
56
57 multiple occasions is unlikely. However, as the tests are designed to measure the same
58
59
60

1
2
3 construct, such as applied medical knowledge for example, the test-retest reliability can be
4
5 incorporated as the correlation between the two test occasions (Test 1 and Test 2 in our
6
7 example). It is important to consider what is being assessed by the tests at each time-point
8
9 when considering use of the RCI. In progress tests, or knowledge tests, where the content
10
11 across all tests is drawn from a pool of 'all knowledge covered by the curriculum', test-retest
12
13 reliability can be used. Where the tests measure different constructs, or only subsets of
14
15 items measure the same construct it may be more appropriate to create subsets of these
16
17 items for analysis of reliable change. Educationally, development in knowledge within a
18
19 domain or topic area is usually of most interest, and it is these instances of retesting
20
21 common constructs to which the RCI can add an additional dimension of understanding. As
22
23 discussed above, for example, identifying particularly excelling students, or identifying
24
25 students who are struggling to develop their knowledge within a particular domain.
26
27
28
29

30
31 Related to the incorporation of the reliability is the calculation of the standard error
32
33 of measurement in the RCI formula presented above (Equation 1), derived from the work of
34
35 Jacobson and Truax (1). In Equation 1, the element incorporating this is:
36
37
38
39
40

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

41
42
43
44
45 However, Maassen (14) highlights methods of calculating this which may be considered less
46
47 reliant on distributional assumptions. As traditional assessment analyses typically rely on
48
49 assumptions such as normality, we have focussed on and presented examples using the
50
51 Jacobson and Truax (1) formulae, but would recommend Maassen's (14) work to the
52
53 interested reader or those who routinely work with skewed data. Similarly, given the RCI's
54
55 focus on change over time, regression to the mean may be an issue. In such cases, we
56
57
58
59
60

1
2
3 suggest the RC_{ID} formula proposed by Hageman and Arrindell (15). The calculation of these
4
5 is more complex, but those wishing to explore the robustness of the RCI in relation to
6
7 multiple test occasions are likely to find their discussions valuable.
8
9

10 11 12 **Usefulness as a research and educational tool** 13

14
15 Most research is conducted at the group level and is focussed on testing hypotheses which
16
17 can be generalised to a wider population, but where the interest is on individuals or smaller
18
19 subgroups the RCI is a useful tool for both research and education. It can be used to
20
21 evaluate interventions designed to improve the learning and experiences of subsets within
22
23 cohorts, for which analysis at an individual level is perhaps more appropriate.
24
25

26
27 Although we would not argue that the RCI is by any means a replacement for group-
28
29 level analysis (e.g. see Alternative Approaches and Limitations section), the RCI also
30
31 provides a useful means of giving individual level feedback to students, especially as it
32
33 overcomes the typically prohibitive demands on resources for producing individualised
34
35 feedback. This may be of particular value in quantitative examinations such as assessments
36
37 over multiple test occasions, or providing feedback on progress from year to year. This is
38
39 particularly true as the approach allows accessible visual feedback and there is body of work
40
41 which suggests that individual feedback in any form is of much more use to the student than
42
43 group-level feedback. Furthermore, this focus on individual change and ability to track and
44
45 quantify individual progress are lacking in more common statistical approaches. The RCI
46
47 therefore allows educators to identify struggling students who may otherwise be
48
49 overlooked if they show improvement, which is not reliable, and hence tailor support tools
50
51 for these students.
52
53
54
55
56
57
58
59
60

1
2
3 When samples or cohorts are very small and group-level analyses are not feasible,
4
5 the RCI provides an alternative approach. It lends itself to research or feedback in situations
6
7 involving small subsets of students (i.e., ethnic sub-groups; individuals with learning
8
9 disabilities), but can still be applied to larger cohorts. As suggested above, this is particularly
10
11 useful when the focus is on specific populations or subgroups which may have smaller
12
13 memberships or be difficult to recruit from. Because of the focus on the individual rather
14
15 than the group as a whole, RCI scores also allow classification of people into those whose
16
17 performance has been reliably altered by an intervention and those whose scores have not,
18
19 as well as the identification of performance profiles which are unusual.
20
21

22
23
24 However, given its dependence on reliability, which may be influenced by a number
25
26 of factors, we would not suggest it be used for high-stakes decisions; its strengths lie in
27
28 supplementing more routine analyses.
29
30

31 32 33 **In Practice**

34
35 Although the RCI provides a useful tool for tracking individual change, in practice, there are
36
37 very few instances where this is the only topic of interest. However, in most research
38
39 designs the RCI can be used alongside traditional tests in order to add a further dimension
40
41 to the findings and our understanding of the data. With respect to progress testing for
42
43 example, changes in performance between two time-points might be visualised using
44
45 scatter plots. This could be augmented by including boundaries for reliable change. In
46
47 addition, the performance of students who reliably improve or deteriorate could
48
49 complement cohort item analysis data in order to provide a clearer indication of which
50
51 groups of students perform well or less well on particular questions, or to inform discussions
52
53 during item review.
54
55
56
57
58
59
60

1
2
3 The use of the RCI to augment more traditional approaches to assessment and
4
5 evaluation, alongside discussion of its uses in the realm of remediation, highlights the need
6
7 to keep in mind the practical application of the RCI, and traditional statistics more generally,
8
9 when making decisions related to student assessment. In particular, regression to the mean
10
11 may explain particularly sudden increases or decreases in student scores. Although this can
12
13 be controlled for to some extent when processing consecutive examinations (16),
14
15 consideration of the reliability of changes over longer periods might help to provide a more
16
17 accurate and robust picture of a student's progression. Furthermore, factors such as low
18
19 scores which need to be discounted due to extenuating circumstances need to be
20
21 thoroughly checked as outlying data could skew the distribution and in turn skew RCI
22
23 calculations.
24
25
26
27

28
29 An additional point to consider in using the RCI to track change over multiple exams
30
31 is practice effects. Although 'practice' is a necessary characteristic in education, to what
32
33 extent similarities between exams might account for performance changes is worth
34
35 considering before drawing conclusions from the results. Although identical exams are
36
37 unlikely to be administered on multiple occasions to the same students, the correction for
38
39 practice effects discussed by Chelune, Maugle, Lüders, Sedlak and Awad (17) and Temkin,
40
41 Heaton, Grant, and Dikmen (18) might be considered if the investigator were interested in
42
43 trying to minimise the influence of potential practice effects or changes in exam strategy
44
45 though they may often be confounded.
46
47
48
49

50 51 52 **Implementation in Excel, R, and STATA** 53 54 55 56 57 58 59 60

1
2
3 As mentioned above, RCI procedures are not included as standard in popular statistical
4 software packages. However, applying the equations provided here in packages such as
5 Excel, R, and STATA is relatively straightforward.
6
7
8
9

10 For presentation to students and staff, the information provided by the RCI and cut-
11 off points is perhaps best presented graphically, as in Figure 1. The package ggplot2 (19) was
12 used in R to create the image for the current paper, but similar results can be achieved with
13 other packages, and with other software. Resources for implementing the RCI using R,
14 STATA, and Excel can be found at <http://tinyurl.com/pppxwr3>, and the authors are more
15 than happy to discuss these resources (contact details above).
16
17
18
19
20
21
22
23

24 25 26 **Alternative Approaches and Limitations**

27
28 As we have noted in this article, the RCI is not a substitute for reliable measurement, which
29 is a particular concern in the measurement of change. What it does do is provide a relatively
30 simple method for accounting for the quality of measures, and uses this information as a
31 tool to help students and help educators improve assessment and practice. The RCI also
32 takes the individual as the unit of interest, supplementing analysis questions which typically
33 assess performance differences at a group level.
34
35
36
37
38
39
40
41
42

43 The RCI is grounded in the assumptions of classical test theory (20) , which is likely
44 the context within which most readers understand issues of measurement reliability and
45 measurement error. Numerous other techniques for assessing change also exist (21). One
46 prominent alternative approach is that of Item Response Theory (IRT; 22); which comprises
47 a framework and set of techniques for dissociating both individuals and items (e.g. test
48 questions) with respect to one or more underlying “latent” dimensions. These approaches
49 show a great deal of merit, though are not without limitations and caveats, such as the
50
51
52
53
54
55
56
57
58
59
60

1
2
3 accessibility of the techniques and statistical software required to implement them. We
4
5 would direct the interested reader to recent reviews on these approaches (e.g. 23).
6

7
8 In spite of the benefits of considering the RCI there are, as with all analyses, some
9
10 caveats to consider. Many of these have already been discussed, but we reiterate the key
11
12 ones here. Firstly, estimates of reliability need to be carefully considered. The RCI was
13
14 developed to incorporate test-retest reliability when evaluating change over time, but given
15
16 the nature of educational assessments, correlations between the two test occasions may be
17
18 the best available indicator of reliability. Secondly, as with typical analyses, the RCI may also
19
20 be influenced by other forms of measurement error, regression to the mean, and practice
21
22 effects. There are various adjustments that have been proposed for accounting for these
23
24 (e.g. reference 15 introduced above), but it is up to the researcher to determine if the
25
26 required data can be obtained, and if the potential reduction in biases outweighs the added
27
28 complexity. Finally, careful consideration of all assessment analyses should be used
29
30 collectively. Information from the individual-level RCI data can inform interpretation of
31
32 group-level analyses and vice-versa to reach defensible and robust conclusions about
33
34 student progression.
35
36
37
38
39
40
41
42

43 **Summary**

44
45 In summary, the reliable change index has a number of potential uses in medical education,
46
47 particularly when the individual is the focus of consideration. It is particularly suited to
48
49 analysis of change over time in individuals, or small subgroups which cannot be captured
50
51 using the more common analysis techniques. In these instances it provides a simple way of
52
53 accounting for the reliability of the measure and can indicate where change is over and
54
55 above the variability of the measurement tool. Although careful thought is needed with
56
57
58
59
60

1
2
3 respect to derivation of parameters needed in its calculation and whether subsequent
4
5 adjustments for other forms of measurement error are needed, it provides a means of
6
7 gaining more from routinely produced assessment data. This can then be used to improve
8
9 and inform decisions relating to student growth, assessment design, and student feedback.
10
11
12
13
14
15

16 **References**

- 17 1. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining
18
19 meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*.
20
21 1991;59(1):12-9.
22
23
- 24 2. Jacobson NS, Roberts LJ, Berns SB, McGlinchey JB. Methods for defining and
25
26 determining the clinical significance of treatment effects: description, application, and
27
28 alternatives. *Journal of Consulting and Clinical Psychology*. 1999;67(3):300-7.
29
30
- 31 3. Zahra D, Hedge C. The reliable change index: Why isn't it more popular in academic
32
33 psychology? *Psychology Postgraduate Affairs Group Quarterly*. 2010;76:14-9.
34
35
- 36 4. Cronbach LJ, Furby L. How we should measure change - or should we? *Psychological*
37
38 *Bulletin*. 1970;74(1):68-80.
39
40
- 41 5. Lord FM. The measurement of growth. *Educational and Psychological Measurement*.
42
43 1956;16(1):421-37.
44
45
- 46 6. Edwards JR. Ten difference score myths. *Organizational Research Methods*.
47
48 2001;4(3):265-87.
49
50
- 51 7. Rogosa D. Myths about longitudinal research. In: Schaie KW, Campbell TR, Meredith
52
53 W, Rawlings SC, editors. *Methodological research in ageing research*. New York: Springer;
54
55 1988. p. 171-210.
56
57
58
59
60

- 1
2
3 8. Rogers PJ. Myths and methods: "Myths about longitudinal research" plus
4 supplemental questions. . In: Gottman JM, editor. The analysis of change. Hillsdale, New
5 Jersey.: Lawrence Erlbaum Associates.; 1995. p. 3-65.
6
7
- 8
9
10 9. Willet JB. Questions and answers in the measurement of change. Review of
11 Educational Research. 1988;15(1):345-422.
12
- 13
14 10. Zimmerman DW, Williams RH. Reliability of gain scores under realistic assumptions
15 about properties of pretest and posttest scores. British Journal of Mathematical and
16 Statistical Psychology. 1998;51(2):343-51.
17
- 18
19 11. Zumbo BD. The simple difference score as an inherently poor measure of change:
20 Some reality, much mythology. In: Thompson B, editor. Advances in Social Science
21 Methodology Greenwich, Connecticut: JAI Press; 1999. p. 269-304.
22
- 23
24 12. Cohen J. Statistical power analysis for the behavioural sciences. 2nd ed. Hillsdale, NJ:
25 Lawrence Erlbaum Associates; 1988.
26
- 27
28 13. Field A. A bluffer's guide to effect sizes. Postgraduate Affairs Group Quarterly.
29 2006;58(March):9-23.
30
- 31
32 14. Maassen GH. The standard error in the Jacobson and Truax reliable change index:
33 the classical approach to the assessment of reliable change. Journal of The International
34 Neuropsychological Society. 2004;10:888-93.
35
- 36
37 15. Hageman WJJ, Arrindell WA. A further refinement of the reliable change (RC) index
38 by improving the pre-post difference score: introducing RC_{ID} . Behavior Research and
39 Therapy. 1993;31(7):693-700.
40
- 41
42 16. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how
43 to deal with it. International Journal of Epidemiology. 2005;34:215-20.
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 17. Chelune GJ, Naugle RI, Lüders H, Sedlak J, Awad IA. Individual change after epilepsy
4 surgery: practice effects and base-rate information. *Neuropsychology*. 1993;7(1):41-52.
5
6
7 18. Temkin NR, Heaton RK, Grant I, Dikmen S. Detecting significant change in
8 neuropsychological test performance: a comparison of four models. *Journal of The*
9
10
11
12
13
14
15 19. Wickham H. *Ggplot2: Elegant graphics for data analysis*. New York: Springer; 2009.
16
17 20. Novick MR. The axioms and principal results of classical test theory. *Journal of*
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
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.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figures

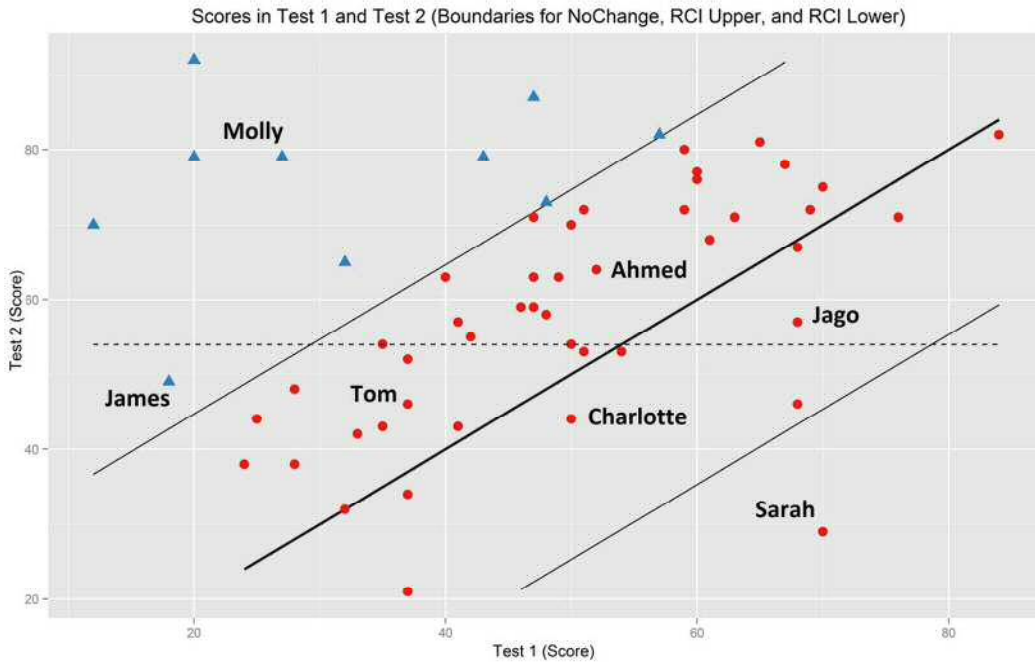


Figure 1: Scatter-plot showing Test 1 and Test 2 scores with a line of no-change (solid diagonal at y=x) upper and lower bounds for reliable change (narrow diagonals) and the mid-point of the means for each test (dashed horizontal)