Manual for Raven's Progressive Matrices and Vocabulary Scales

By

J Raven, J C Raven and J H Court

Section 3

Standard Progressive Matrices

(including the Parallel and *Plus* versions)

2000 Edition

With norms for the SPM *Plus* and formulae for calculating change scores

VALUE OF CHANGE SCORES

Important Note:
This material forms part of the above manual and should only be used in conjunction with its General Introductory Section (General Overview).

The value of change scores and the measurement of change in groups and individuals[253]

As we have seen, the concept of "learning potential", to be measured from the difference between people's scores before and after a period of training, became popular in the last quarter of the previous century. While variants of this methodology – especially those of Guthke – still hold out great promise, enthusiasm declined as it became evident that it was, in some sense, "easier" to achieve a "gain score" of, for example, "5" from, say, a raw score of 15 to 20 than from 55 to 60. This arose in part from the uneven distribution of differences between item difficulties discussed earlier. This problem, taken together with other problems discussed in Prieler and Raven[254], meant that such "gain scores" had different meanings at different parts of the scale and thus little predictive validity.

But "learning potential" is not the only psychological characteristic which researchers would like to assess from the difference between an individual's pre- and post-test scores. It is a common observation that, for example, some people are much more affected by stress than others. Could individual differences in such susceptibility be measured by the difference between their scores before and after a stress-inducing event? Then again, regardless of their absolute score, do some people react more – ie do their scores change more – than others to certain types of stress (or drug), and does the magnitude of that reaction have predictive validity?

A statistically related clinical question is "relatively *how much*" does an *individual* patient improve or deteriorate in response to different regimens? Does he *really* improve when his score goes from A to B, or are the items at that point in the scale too closely spaced for the effect to be worth noting? Is a noteworthy effect of one treatment followed by an even more marked effect of another?

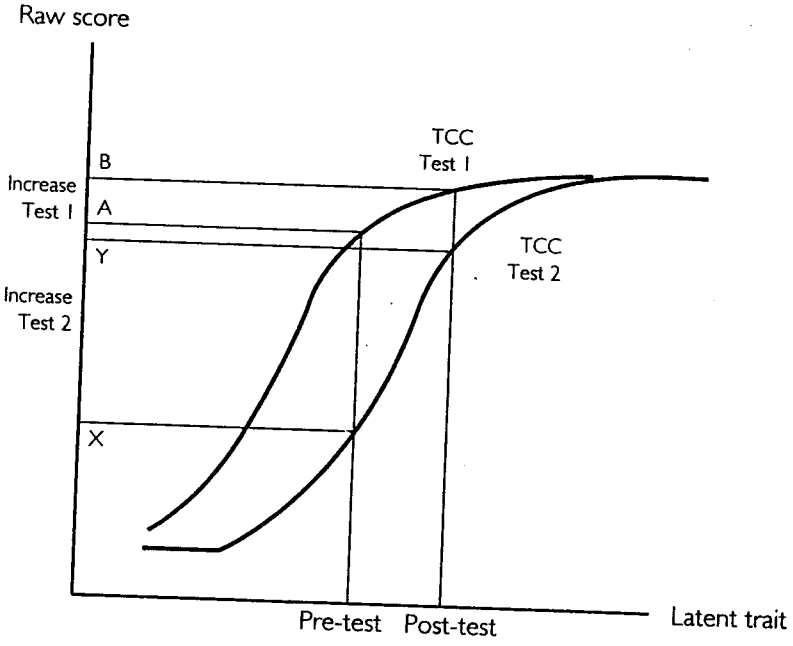
A similarly statistically related, but this time conceptually very different, problem has to do with the measurement of the *differential* effect on, for example, more and less able *groups* of people of such things as experimental laboratory manipulations or educational or social programmes. We have seen that it has been common to claim that the changes over time discussed earlier have been "greater" among the less able than the more able. Such claims can have little meaning if they are based on raw scores and the differences between the item difficulties are different at different points in the scale (for example, smaller at the bottom end than at the top end), and if, in particular, the test used (such as the Classic SPM) has a ceiling which does not permit those with high scores to demonstrate gains which have, in reality, occurred.

These problems have been fully discussed by Prieler and Raven[255]. But the most disturbing observation that has emerged from this work is that, as Fischer[256] has shown, *the apparent relative gains of high and low scoring respondents is heavily dependent on the absolute difficulty of the test used* ***even if the test used conforms to the Rasch model.***

The problem may be indicated as follows.

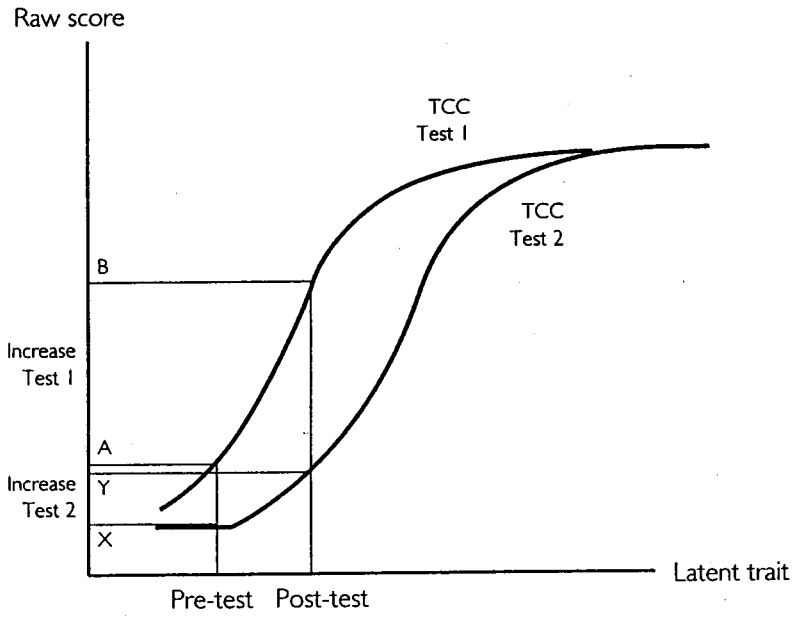
Figure SPM7 illustrates the problem of determining gain among high ability respondents in a way which would make it possible to meaningfully compare that gain with the gain of those with average or low scores. Figure 8 provides a parallel illustration for people of low and average ability.

Figure SPM 7
Illustration of changes in raw scores on "Easy" and "Difficult" IRT-based tests of cognitive ability for identical changes in latent ability.
***High* ability group only**



If we employ a test (Test 1) having the Test Characteristic Curve shown on the left in Figure SPM7, the mean scores of the high ability group increase from A at the pre-intervention level (ie before training or the administration of a drug) to B after the intervention. This is a relatively small increase. But if we use the more difficult test (Test 2), whose TCC is shown on the right, the same increase in the latent ability of the high ability group shows up as a *huge* increase in raw score, moving from X to Y.

Figure SPM 8
Illustration of changes in raw scores on "Easy" and "Difficult" IRT-based tests of cognitive ability for identical changes in latent ability.
***Low* ability group only**



As can be seen from Figure SPM8, exactly the opposite effect occurs at the other end of the scale. The apparent increase in score from pre-test to post-test is huge on Test 1 and trivial on Test 2.

Putting the two cases together, it is obvious that, if a researcher employs Test 1 to assess the impact of an intervention (or to assess the effects of passage of time), the relative gains of the low ability group are huge while those of the high ability group are trivial. On the other hand, if the researcher employs Test 2, exactly the opposite findings emerge.

The general, and vitally important, conclusion illustrated using these examples is that the apparent magnitude of any real increase in latent ability arising from an intervention (such as a developmental experience, a stress-inducing event, or administration of a drug), or from a natural change over time, depends on (a) the difficulty level of the test used, (b) the shape of its Test Characteristic Curve, and (c) the sector of the curve on which the change occurs.

This makes it virtually impossible, without employing techniques like those developed by Fischer and Prieler (and described in Appendix SPM3), to (1) make any meaningful statement about the *relative* magnitude of gains or losses

of high, medium, and low ability groups, (2) generate meaningful individual change scores (ie to measure individual differences in learning potential or responsiveness to stress), or (3) to decide whether a change in an individual's score represents an improvement or deterioration worthy of note.

The basis of the methodology which makes it possible to handle these problems is briefly described in Appendix SPM3. Here it is sufficient to illustrate the principle on which it is based. When the same item – even in a non-Rasch-homogeneous test – has been administered at two points in time, the two administrations can be viewed as if they constituted a 'miniature' Rasch scale. Since the same *item* has been administered at two points in time the "Scale" must measure the same latent trait. Such a "scale" must therefore be Rasch homogeneous. In effect, one has as many miniature Rasch scales as there are items in the test. For example, if one presents the same 10 items at pre-test and post-test, one has 10 miniature Rasch scales. There is no requirement that these items *as a group* tap any common dimension. They could, indeed, be, and sometimes are, actively chosen to measure 10 *different* dimensions so as to get the maximum information in the minimum amount of time. Interestingly, however, in a second step, one can test if any effects detected are general across all items. If this is the case (and, from the many studies available, it would seem that it usually is), one can calculate the overall effect size of the treatment. This may at first seem strange, but the advantage is that it is very flexible. Of course, if the test used is in fact unidimensional the procedure is unnecessarily complicated. However, in clinical studies, it is common to use multidimensional questionnaires or more than one test covering several dimensions. The relative effect size on the different components can then be assessed.

The same procedures can be applied to identify which *people* have changed.

When it is desired to assess changes *within individuals* over time or in response to an intervention, or when one wants to compare one individual with another, it *is* necessary to employ a Rasch homogeneous test. Nevertheless, the effect sizes calculated are again independent of the distribution or content of the items used, the difficulty of the test, and the shape of its Test Characteristic Curve.

A nomogram for determining the relative importance of deteriorations or improvements in SPM *Plus* scores within individuals for people of different levels of ability is included in Appendix SPM 3.

Appendix 3: An IRT-Based Methodology for the Assessment of Change

By Gerhard H. Fischer and Jörg A. Prieler

In this Appendix we will attempt to outline, in a non-technical language, the basis of the IRT methodology which has been developed to overcome the problems posed by the fact, discussed earlier, that the apparent magnitude of changes in the scores of groups or individuals depends on (a) the general difficulty level of the test relative to the ability tested and (b) the distribution of the item parameters relative to the interval on the latent trait where change occurs.

The solution of these problems depends, like the *Progressive Matrices* itself, on the deployment of Item Response Theory (IRT), a brief introduction to which is found in the *General Introductory Section* of this *Manual*. Nevertheless, despite the fact that development of the methodology depended on IRT, its application is not strictly limited to IRT-based tests. It can also in many instances be used to obtain meaningful measures of change employing tests constructed according to classical test theory[309].

There are two main areas of application of the IRT methodology for measuring change:

(1) The measurement and statistical assessment of change in *groups* (a) over time, (b) in response to different types or dosages of treatment(s), (c) in response to the same treatment(s) at different levels of ability, (d) differing in personality traits, gender, age, or any other observable characteristics; and

(2) The measurement and statistical assessment of change in *individuals* (a) over time, (b) in response to different types or dosages of treatment, and (c) in response to the same treatment(s), irrespective of the individual level of ability.

The measurement and statistical assessment of change in *groups*

It is easiest to illustrate the principle by discussing a situation in which it is desired to document the differential effect of an experimental treatment (such as an educational enrichment programme) on high and low ability respondents although, as we shall shortly see, application of the method is by no means limited to such situations.

When the same test – even a test that is not unidimensional – has been employed to assess performance before and after an intervention, each *item* that has been presented on the two occasions can be treated as if it were a pair of items with **different** item parameters within a common Rasch scale, that is, as if it were a 'miniature Rasch scale' of length 2. For example, if one presents the same 10 items as pre-test and post-test, one thereby obtains 10 miniature Rasch scales. There is no requirement that these items measure a common dimension; they could indeed be, and, in clinical studies often are, actively chosen to measure 10 *different* dimensions in order to monitor change as *comprehensively* as possible.

Despite this, there is no need to use long tests, because each item measures a different latent dimension. (These dimensions may be correlated, or in some other way mutually dependent, or independent.)

After this, one can, *in a second step*, assess whether any effects detected generalise across all items. If they do (and, from the many studies available, it would seem that it is indeed often the case), one can estimate an overall effect size for the treatment(s), or otherwise assess the relative effect sizes on the different "dimensions" involved. Obviously, the result is a very flexible set of procedures.

Although the development of these procedures is formally grounded on IRT, the method mentioned departs fundamentally from the unidimensionality assumption of most IRT models, For this reason the present model of change has been termed the "Linear Logistic Model with Relaxed Assumptions" (LLRA)[310]. It is implemented in the software LPCM-Win 1.0[311].

Variations and extensions of the method to (a) tests that are known to be unidimensional and (b) to items with more than two ordered response categories, are also available. (The reader interested in the psychometric background of this approach is referred to the book *Rasch Models* edited by Fischer & Molenaar[312], and to the *Handbook of the Usage of LPCM Win 1.0*[313].

The LLRA and the other related models for the measurement of change allow various types of designs of studies:

(i) Presentation of the same item sets at two or more time points to the same person groups. The items may, but need not be, unidimensional.

(ii) Presentation of different, possibly overlapping, item samples from a unidimensional item pool, at two or more time points. One or more unidimensional item pools may be used within the same study, so that the total item samples again become multidimensional. In such cases it is important that, at each time point, at least one item is selected from each unidimensional item pool, assuring that the respective latent dimensions are actually measured at each time point. In principle, there is no limitation to the number of latent dimensions that can be included, except for obvious practical limitations of test length.

(iii) The items may be dichotomous (as in most ability tests) or polytomous (with ordered response categories, as in many clinical rating scales).

(iv) There may be any number of treatment and control groups. A treatment group is, by definition, a group of persons responding to the same subsets of items at the same time points and receiving the same treatments or treatment combinations.

(v) The data may be complete or incomplete. Incompleteness of the data (eg, missing observations) entails, however, that formally the number of treatment groups increases, because all persons within one treatment group must have responded to the same item subset.

Obviously, these features of admissible research designs cover a wide range of possible studies. Given that a study is designed meaningfully with respect to the realised treatment combinations, the application of the IRT methodology mentioned will yield estimates of effect parameters of the treatments as well as of one or several trend effects representing causes of change unrelated to the treatment(s). The method also yields significance tests and standard errors for the effect parameters. Moreover, LPCM-Win supports the formulation and testing of a number of standard hypotheses (eg, generalisability of treatment effects or of amounts of change over [a] items subsets and [b] person subgroups) as well as of a host of customised hypotheses.

To make some of these advantages more concrete, consider once again a study in which the differential impact on different types of children in a pre-school education programme is to be assessed. In this case, most of the participants will, at the time of the post-test, achieve higher than even the most able ones did at the time of the pre-test. Consequently, it is necessary to present, for the post-test, more difficult items than those used for the pre-test.

In these cases, it is possible to use the so-called Hybrid LLRA. This combines Rasch-homogeneous item pairs with the multidimensionality of LLRA. It depends on having found, however, via an item calibration study, Rasch-homogeneous item pairs. The easier item out of each pair is presented at the pre-test, and the more difficult at the post-test. In this way it is possible to ensure that the post-test items will be of an adequate difficulty level. One possible variant of this would be to use a subset of items out of a Rasch-homogeneous test (such as the SPM *Plus*) at the pre-test, and the others, appropriately selected, at the post-test.

At this point, because use will be made of it later, it is desirable to explain the conceptual shift that makes it possible to use IRT to solve these hitherto intractable problems in the measurement of change. In essence, one fundamental knack is to use a shift in *item* parameters (which generated the miniature Rasch scales in our previous discussion) as an index of change within *persons*. Technically speaking, the same item presented to respondents at two time points is formally considered as a pair of "virtual" items with different item parameters. The difference between the item parameters within pairs becomes an indicator of change in the respondents on the respective latent dimension. Under the assumption of generalisability of change over the latent dimensions measured by different items and over persons within a treatment group, each pair of virtual items contributes to the overall information on the amount of change in that group. Therefore, combining all these contributions enables a measurement and statistical evaluation of change.

The estimation of effect parameters using the LLRA does not involve the estimation of item or person parameters. Only *change* parameters (ie the effects of treatment or changes which have occurred over time) are estimated. The computation is based entirely on those response combinations where a person has solved *only one* of the items of an item pair (=miniature Rasch scale). Response combinations where both responses to the items of a pair have been correct or both incorrect, provide no information on change and have to be ignored. That is, it is advantageous to maximise the numbers of scores 1 (and neither 0 nor 2) on each of these miniature Rasch scales (item pairs). This can be achieved by an intelligent selection of the items in forming the pairs mentioned.

The measurement and statistical assessment of change in *individuals*

A different research motivation leads to the study of individual change: clinical psychologists, for instance, ask whether a patient has been able, after a treatment period, to improve his/her test performance level; educational psychologists want to compare individual growth within a certain time period to the average growth of the cohort; applied psychologists are interested in the amount of change on a trait in an individual after some training or personality development programme, etc. The tests used are sometimes achievement tests with dichotomous items (such as the SPM *Plus*), sometimes scales composed of items with several (ordered) response categories, like "always", "mostly", "rarely", "never". For all these cases, an IRT methodology is now also available that makes it possible to measure individual change and to evaluate change statistically.

It has to be stressed, however, that - unlike in the case of group-oriented studies - the item pool used must be unidimensional. This is not hard to explain, though: if a study focuses on individuals, and if each item possibly measures a different dimension, only two discrete responses are available per latent dimension, rendering a scientific result on the amount of change on each latent continuum impossible. Besides the restriction of the unidimensionality of the item pool, the present methodology has so far been developed only for two time points. In studies with more than two time points, pairs of them must be analysed separately.

On the other hand, there is a great flexibility with respect to the composition of the tests used at the two time points: from the given unidimensional item pool, for each time point any subset of items may be selected. Therefore, the same items may be given to the respondent twice, or entirely different subsets of items may be selected for the pre- and post-test, respectively, or the two items sets may overlap partially. So it is up to the researcher to select the items according to his/her theory or purpose. If the researcher expects, for instance, an increase in the respondent's score on the ability or trait measured, he/she may choose easier items for the pre-test than for the post-test, so that the expected shift on the latent dimension is roughly compensated for by an increase of item difficulty.

The idea behind the psychometric method is that the amount of change in the studied individual is projected into the item parameters: the concept of "virtual" items again turns out to be essential for understanding the approach. Instead of thinking in terms of a change of the person (ability) parameter, it is helpful to imagine change as a shift of the post-test item parameters relative to the pre-test item parameters. Therefore, the person (ability) parameter - in spite of its change in reality - is technically

considered as a constant, while the item parameters of the post-test items are exchanged by virtual item parameters. As a consequence, the responses given by the individual on both tests can be treated like responses of a respondent to just one test, the length of which is the sum of the lengths of the pre-test and of the post-test. This makes it possible to employ the so-called "conditional maximum likelihood method". Its advantage is that the person parameter is eliminated from the further steps in the estimation and statistical testing procedures. Moreover, this approach avoids any asymptotic approximations since only the exact conditional distribution of the gain score is used.

This is not the place to outline the methodology in any formal detail[314]. It is sufficient to say that the method yields, for each individual, an estimate of the amount of change on the latent dimension, that this measure is independent of the true initial level of the trait or ability, that confidence intervals can be computed for the true individual amount of change, and that the amount of change can be tested for significance. A good example of a practical application of the methodology can be found in Prieler[315]. In this study, a comprehensive battery of tests was administered to officer cadets before and after a strenuous night march in order to determine which *change* scores best predicted aptitude for the intended career.

Confidence intervals for the change parameter and significance tests can be obtained in either of two ways. A simple and straightforward approach is to compute so-called Clopper-Pearson confidence intervals and related significance levels. This method has several attractive properties: there is a certain "double monotonicity" in the output tables corresponding with what one would expect from a substantive point of view; the results are unique in the sense that, for given raw scores of an individual, they are uniquely determined (except for certain "boundary cases" which are of little substantive interest anyway). For tests with dichotomous items conforming to the Rasch Model, these Clopper-Pearson intervals and related levels of significance can be obtained by means of LPCM-Win 1.0 (see above). For scales composed of polytomous ordered response items, however, methods recently described in Fischer[316] are needed. Tables SPM43 through SPM46 have been computed by means of a so far unpublished software of the first author.

The output of this software comprises individual change parameter estimates, confidence intervals for the true amount of change, and levels of significance of the observed change. To illustrate some of the possible uses of this methodology, we include Tables SPM43 and SPM44. In the leftmost column, Table SPM43 gives all possible raw scores attainable in the SPM *Plus* test, ranging from 0 to 60, on the assumption that the complete test had been given as the pre-test (these raw scores are denoted by r_1; for convenience, the same raw scores are repeated in the rightmost column). Similarly, the top row gives all possible post-test scores, denoted by r_2, assuming again that the entire test has beeen administered. (To ease the reading of the table, the raw scores r_2 are repeated in the bottom row). The interior of the table contains the levels of significance of the difference between post-test score and the pre-test score, supposing that the respective raw score in the pre-test was r_1. A dot represents significance at the .10 level, an "s" significance at the .05 level, an "S" significance at the .01 level, and a "T" significance at the .001 level.

It will be seen that, over a wide range of raw scores, a score difference $r_2 - r_1$ of app. 8 means a significant increase, and a score difference $r_2 - r_1$ of app. -8 a significant decrease of test achievement (at a significance level of .05). Table SPM43 is based on a two-sided statistical test procedure. Notice that on both sides of the main diagonal the significance levels decrease monotonely, both in the horizontal and vertical directions. This is what is meant by "double monotonicity".

Another way of presenting the same information is shown in Table SPM44. The leftmost column gives the pre-test score r_1 of an individual, the other columns those post-test scores r_2 which depart significantly from the given pre-test score r_1, for significance levels ranging from 0.1 to .001. The symbol "ns" indicates that, for these post-test scores, the amount of change is not significant.

The disadvantage of this method is that the confidence intervals and the related significance tests are somewhat "conservative". This term from statistics means that significance is sometimes not attained,

although a more powerful test would yield significance. Fortunately, techniques exist for the construction of "uniformly most accurate" (UMA) confidence intervals and "uniformly most powerful" (UMP) tests. The disadvantage of both of them is that they are grounded on so-called "randomised" scores. Randomisation in this context means that a small continuous random component is added to the observed discrete gain score, transforming the latter into a continuous random variable. For readers who care about the formal details: this random component is independent of the gain score and has a rectangular distribution on the interval [0,1). It is counter-intuitive, but true, that the addition of that random component leads to an increase of the precision of the confidence interval for the change parameter and to an increase of the power of the statistical test.

One problem of this approach is obvious: since an independent random component has to be added to the gain score in each individual case, repeating the procedure for two persons with exactly the same raw scores in both the pre-test and post-test, may occasionally lead to a different assessment of the significance! This further implies that the "double monotonicity" of the output tables is lost, which may be considered counter-intuitive by some users. Finally, no fixed tables like Tables SPM43 or SPM44 can be printed, because the outcome depends in part on the random component.

What we can do, however, is print a table where we fix the random component at a value of .5. The intuitive idea behind this is that .5 is the median value of the random component, so that the boundaries of the significance levels in the table are the median boundaries; this means that a slightly narrower confidence region (or boundaries of the significance levels) is to be expected in 50% of all cases, and a wider confidence region in the other 50%. The results of this are shown in Tables SPM45 and SPM46, which otherwise are completely analogous to Tables SPM43 and SPM44, respectively. It is obvious that, by means of the randomisation, more precise results about the change parameter and its significance are obtained: while in Table SPM43 over a wide range of pre-test scores, an increase by approximately 8 points was significant at the .05 level, an increase by 7 points is now sometimes equally significant.

Limitation of space prohibits giving further details about this new methodology. A much more detailed – but also much more formal – account of its possibilities can be found in Fischer[317].