Controlling response dependence in the measurement of change using the Rasch model

Abstract

The advantages of using person location estimates from the Rasch model over raw scores for the measurement of change using a common test includes the linearization of scores and the automatic handling of statistical properties of repeated measurements. However, the application of the model requires that the responses to the items are statistically independent in the sense that the specific responses to the items on the first time of testing do not affect the responses at a second time. This requirement implies that the responses to the items at both times of assessment are governed only by the invariant location parameters of the items at the two times of testing and the location parameters of each person each time. A specific form of dependence that is pertinent when the same items are used is when the observed response to an item at the second time of testing is affected by the response to the same item at the first time, a form of dependence which has been referred to as *response dependence*. This paper presents the logic of applying the Rasch model to quantify, control and remove the effect of response dependence in the measurement of change when the same items are used on two occasions. The logic is illustrated with four sets of simulation studies with dichotomous items and with a small example of real data. It is shown that the presence of response dependence can reduce the evidence of change, a reduction which may impact interpretations at the individual, research, and policy levels.

Key words

Rasch model, local independence, response dependence, measurement of change

David Andrich

Graduate School of Education, The University of Western Australia,

35 Stirling Highway, Crawley, Western AUSTRALIA, 6009

Telephone: +61 8 6488 1085; Fax: +61 8 6488 1052

Email. david.andrich@uwa.edu.au

Controlling response dependence in the measurement of change using the Rasch model

1. Introduction

The measurement of change at both the group and individual level has generated a large literature. The mainstreaming of modern test theory has overcome a number of concerns including those which arise from the use of raw scores. However, in the measurement of change when the same instrument is used on more than one occasion concern has been raised that responses on the second occasion will be similar to responses on the first occasion simply because the same items are responded to even if the property being assessed has changed. For example, even though a respondent's proficiency or attitude might have changed, the response on the second occasion will be more similar to the first occasion than it should be given the change. Clearly, the response dependence of this kind will bias the person estimates on the second occasion giving rise to misinterpretation of actual changes due to treatments or some other reasons. The purpose of this paper is to explain this kind of response dependence between two occasions and to present a mechanism for assessing, controlling, and eliminating the bias it produces with the application of the Rasch measurement theory (RMT) of modern test theory.

The Rasch models have some advantages over other models, including those concerned with total raw scores obtained by summing the integer responses to items. The concern with the use of raw scores is that they do not provide interval level measurement, that is, that the same numerical change using observed scores at different parts of the continuum do not correspond to the same quantitative change on the inferred continuum of the variable.¹ It has been argued that Rasch measurement theory (RMT)² overcomes the problem by linearizing the raw scores and that it provides interval level measurement.³ In RMT the response of each person to each item is formalised as a function of a single person parameter and one or more item parameters where the person parameter characterises the property on the continuum as a real number.

This paper is concerned with the application of RMT in the measurement of change which has some advantage over other methods, providing of course that the data fit adequately to the Rasch model used for the analysis. The advantages arise from the sufficiency of the total score for the person parameter, which can be exploited in the estimation of the item parameters. Specifically, first, the person location estimates are obtained by effectively linearizing the total observed scores thus providing interval scale estimates;⁴ second, the estimates of the item parameters can be obtained by conditioning out the person parameters and therefore, if the persons change only in terms of their location the item parameters will remain invariant over the times of assessment;⁵ third, there is no need to make assumptions about the distribution of the persons, for example that the distribution is normal, at either time of assessment. Whether item parameters remain invariant is an empirical question that can be assessed using a variety of methods including those described in terms of differential item functioning (DIF). In addition, if the item parameter estimates are based on conditional methods then they are consistent in the sense that as the number of persons in the data increases, the item estimates converge to their true value.^{6,7}

The application of modern test theory in general and RMT in particular has brought into focus the assumption or requirement of statistical independence among responses, which although generally not made explicit, is also required in classical test theory (CTT).⁸ Following Lazarsfeld and Henry,⁹ the implied statistical independence is referred to as *local independence*. For a group of persons the responses among items will be dependent as a function of the person parameters on the variable, and therefore correlated in a sample. The term *local independence* arises from the property that for persons with the same location parameter, the responses to the items will be uncorrelated. However, the independence is general in the sense that the probability of the specific set of responses of different persons to the same item, or of the same person to different items, is the product of the individual probabilities.¹⁰ Marais and Andrich¹¹ formalised two kinds of violation of statistical independence in the unidimensional dichotomous Rasch model (RM) – first a violation of unidimensionality, second a

violation of response independence whereby the response to an item depends in part on the response to a previous item.

The concept of response dependence in which the response to one item depends in part on the response to a previous item can be generalized to the concept in which the response to an item at a second time depends on the response to the *same* item at the first time. This paper is confined to the case of responses to the same items at two times though the principles are readily generalized to more than two times. Response dependence from a first to a second time can arise because of some idiosyncratic effect of the item which governs the response at both times or where some effect such as memory, which Kuncil and Fiske refer to as *inappropriate components*, affects the response the second time.¹² It is potentially likely to occur in the assessment of health outcomes where change from some treatment is expected and in which the same items are used on multiple occasions. It may also occur in educational and psychological assessment in the same circumstances.

This paper shows how response dependence can be controlled in the sense that, if present, it can be quantified and eliminated in the assessment of change using the RM. The paper exploits the formalisation of response dependence between two items in Marais and Andrich¹¹ and the estimation of the degree of response dependence between them derived in Andrich and Kreiner¹³. Furthermore, it shows that the logic for the elimination of response dependence is remarkably elegant and simple.

2. The Rasch model and the formalisation of response dependence

Because the formalisation of response dependence has been detailed in Marais and Andrich¹¹ and Andrich and Kreiner¹³ only a summary is provided here. First, the unidimensional RM takes the form

$$\Pr\{X_{ni}=x\}=\exp(x(\beta_n - \delta_i))/\gamma_{ni}, \quad (1)$$

where β_n , δ_i are respectively the person and item location parameters, termed respectively proficiency and difficulty in the case of performance assessments which can occur in health outcomes, education, psychology and other areas, X_{ni} is the random variable which takes on the values $x=0,1$ when person n responds to item i and $\gamma_{ni}=1+\exp(\beta_n-\delta_i)$ is a normalising factor ensuring that $\sum_{x=0}^1 \Pr\{X_{ni}=x\}=1$. Clearly the greater the value of β_n relative to δ_i , the more likely the response $x=1$, generally referred to as the positive or successful response. The left panel of Figure 1 shows the probability of a positive response as a function of a person location β for an item with $\delta=0.0$. The probability curves in the right panel of Figure 1 are explained later.

INSERT FIGURE 1

Figure 1 Probability curve for a dichotomous item, called Item 3, in the RM (left panel) and its curves when resolved because of local dependence (right panel).

Second, the principle of statistical independence for the matrix of responses $[x]$, $n=1,2,\dots,N$; $i=1,2,\dots,I$, is expressed as

$$\Pr\{[x]\} = \prod_{n=1}^N \prod_{i=1}^I \Pr\{x_{ni}\}. \quad (2)$$

For the responses of person n to two items i, j , Equation 2 specializes to

$$\Pr\{(x_{ni}, x_{nj})\} = \Pr\{x_{ni}\} \Pr\{x_{nj}\}, \quad (3)$$

from which it follows that

$$\Pr\{(x_{nj} | x_{ni})\} = \Pr\{x_{nj}\}, \quad (4)$$

that is, that the response x_{nj} is not affected by the specific response x_{ni} . Equations 1 and 4 together show that, in addition to being governed by their respective parameters δ_i, δ_j , the probabilities of responses to both items are governed by the same person parameter β_n . Therefore, a large positive value of β_n is likely to produce a positive response to both items and a large, in magnitude, but negative value is likely to produce a negative response to both items. In a sample of persons with variable values of β , there will be a positive correlation between the responses to the items. However, there is no further relationship between the responses, other than that governed by β . Therefore for the same person, or for a hypothetical group of persons with exactly the same value of β , there will be no correlation among the responses between items. These relationships explain formally the use of the term *local independence*.

In the estimation of the item parameters in the RM in which the person parameter β_n is conditioned out and plays no role, a constraint on the estimates is required, and is generally taken to be $\sum_{i=1}^I \hat{\delta}_i = 0$. The person parameters are then estimated taking the item parameters as known, again generally using maximum likelihood (ML) methods. In case the origin in an analysis is required to be the same as a previously set origin, some other constraint might be imposed including using estimates of one or more items from a previous analysis.

2.1 Construction of local response dependence

If the response to item i affects the response to item j then local dependence, referred to as response dependence, violates Equation 3. This violation implies that the response x_{nj} is not only a function of the parameter β_n and the item parameter δ_j , but is also a function of the response x_{ni} . Marais and Andrich¹¹ formalised this response dependence in the form

$$\begin{aligned}\Pr\{X_{nj} = 1 \mid X_{ni} = 1\} &= (\exp(\beta_n - (\delta_j - d)) / (1 + \exp(\beta_n - (\delta_j - d)))) \\ \Pr\{X_{nj} = 1 \mid X_{ni} = 0\} &= (\exp(\beta_n - (\delta_j + d)) / (1 + \exp(\beta_n - (\delta_j + d))))\end{aligned}\quad (5)$$

where $|d|$ is the magnitude of the dependence. The effect d can be taken as positive or negative. However, to reflect that the likelihood that response x_{nj} will be the same as response x_{ni} is greater than it would be without dependence, it is taken that $d > 0$.

Clearly, from Equation 5,

$$\begin{aligned}\Pr\{X_{nj} = 0 \mid X_{ni} = 1\} &= 1 - (\exp(\beta_n - (\delta_j - d)) / (1 + \exp(\beta_n - (\delta_j - d)))) \\ \Pr\{X_{nj} = 0 \mid X_{ni} = 0\} &= 1 - (\exp(\beta_n - (\delta_j + d)) / (1 + \exp(\beta_n - (\delta_j + d))))\end{aligned}\quad (6)$$

Marais and Andrich¹¹ also show that the response space of the four possible responses defined in Equations 5 and 6 form a sample space in the sense that their probabilities sum to 1.

Equation 5, in which it is taken that $d > 0$, clearly violates Equation 4 and is interpreted in the following way. First, suppose again that the assessment is of proficiency of some performance and therefore that δ can be referred to as a difficulty. Then the first line of Equation 5 formalises that when $d > 0$, a correct response $x_{ni} = 1$ on item i reduces the difficulty of the dependent item j from δ_j to $\delta_j - d$, making item j easier than it would be otherwise. This reduction in difficulty increases the probability that the response x_{nj} will also be correct. It is stressed that making the item *easier in the equation* is a mechanism for formalising a statistical effect that the response is more likely to be positive than it would be otherwise, and that the item as such has not become easier for the person. Analogously, the second line of Equation 5 formalises that an incorrect response $x_{ni} = 0$ on item i increases the difficulty of dependent item j from δ_j to $\delta_j + d$, making item j more difficult for the person than it would be otherwise. This increase in difficulty increases the probability that the response x_{nj} will also be incorrect. Again, the item is not literally more difficult; instead it is an algebraic

mechanism that ensures that the probability of a negative response in this case is more likely than it would be otherwise. The two lines of Equation 6 are interpreted in a complementary way. The right panel of Figure 1 shows the probability curves of the resolved items for the item in the left panel of the Figure when $d = 2$.

2.2 Estimating the response dependence effect d of Equation 5

The logic of controlling for possible response dependence when the same items are used at two times arises from the logic of estimating the response dependence effect d of Equation 5 between two items. Therefore, this logic, described in detail in Andrich and Kreiner,¹³ is summarised below.

First, the response x_{nj} to dependent item j is resolved into two items. The first resolved item is composed of responses of those persons whose response on the independent item i is $x_{ni} = 0$. The second resolved item is composed of the responses of those persons whose response on independent item i is $x_{ni} = 1$. Let these resolved items be denoted j_{i0} and j_{i1} respectively, and their responses be denoted $x_{nj.i0}$ and $x_{nj.i1}$.

Second, because of inherent, structural dependence of responses between the two resolved items and the original responses, the original responses x_{nj} are removed from the data matrix. Third, because of an analogous structural dependence between the two resolved items and the responses x_{ni} to independent item i , the responses x_{ni} are also removed from the data matrix.

The resultant matrix in which the responses are statistically independent, necessarily has structurally missing data. However, this is generally no impediment to estimation using

modern software. Andrich and Kreiner¹³ showed that if the remaining items, including unresolved ones, conformed to the Rasch model then the response dependence effect d , and its standard error, could be estimated readily. The estimate of d is obtained from the difficulty estimates of items j_{i0} and j_{i1} . Let their respective difficulties be $\delta_{j.i0}$ and $\delta_{j.i1}$. Then from Equation 5,

$$\delta_{j.i0} = \delta_j + d; \delta_{j.i1} = \delta_j - d, \quad (7)$$

giving

$$d = (\delta_{j.i0} - \delta_{j.i1}) / 2. \quad (8)$$

Thus from the estimates $\hat{\delta}_{ji0}$ and $\hat{\delta}_{ji1}$, and their standard errors, $\hat{\sigma}_{ji0}$, $\hat{\sigma}_{ji1}$,

$$\hat{d} = (\hat{\delta}_{j.i0} - \hat{\delta}_{j.i1}) / 2, \quad (9)$$

with a standard error $\hat{\sigma}_d$ given by

$$\hat{\sigma}_d = \sqrt{(\hat{\sigma}_{j.i0}^2 + \hat{\sigma}_{j.i1}^2) / 4}. \quad (10)$$

Andrich and Kreiner¹³ suggested that this procedure for assessing response dependence had the advantage over some other methods, such as correlational ones, in that it quantified the impact on the dependent item's difficulty in terms of the metric of the items. It is this feature which is effectively exploited in the method described in this paper.

3. Accounting for response dependence between the first and second assessments

The resolution of items that permits the estimation of the response dependence effect in the metric of the items can be exploited, perhaps surprisingly, to eliminate the effect in the person estimates. As a result, estimates of person locations at the second time of

testing are unbiased by the response dependence. As with the estimation of the magnitude of response dependence, accounting for response dependence rests on having a formalisation of the form specified in Equation 5.

3.1 Estimating the dependence from assessments at two times on the same items

Let a hypothesised response dependence effect for item k between Times 1 and 2 be d_k . To estimate d_k , the matrix of responses to be analysed includes responses to all items at Time 2, and *in addition*, the responses to item k at Time 1. Let the responses to item k at the two times be $x_{n(k1)}$, $x_{n(k2)}$ respectively. Using the earlier notation for a resolved item, the resolution of $x_{n(k2)}$ gives items $(k2)_{10}$, $(k2)_{11}$ with difficulties $\delta_{k2.10}$, $\delta_{k2.11}$ for respective responses $x_{n(k1)} = 0, 1$ at Time 1. The analysis of this matrix provides estimates of difficulties

$$\hat{\delta}_{k2.10} = \hat{\delta}_k + \hat{d}_k, \quad \hat{\delta}_{k2.11} = \hat{\delta}_k - \hat{d}_k \quad (10a)$$

from which the estimate of d_k , according to Equation 9, is given by

$$\hat{d}_k = (\hat{\delta}_{k2.10} - \hat{\delta}_{k2.11})/2, \quad (10b)$$

with a standard error of

$$\sigma_{\hat{d}_k} = \sqrt{(\hat{\sigma}_{k2.10}^2 + \hat{\sigma}_{k2.11}^2)/4}. \quad (10c)$$

3.2 The maximum likelihood estimate for person locations

To show the effect of response dependence in the estimates of person proficiencies at Time 2 when responses to item k are resolved in the above analysis, it is instructive to review the ML estimate of the person parameter from responses to a fixed set of items.

In the reasoning below it is assumed that the item parameters have been estimated separately by exploiting the sufficiency property of the total score for the person parameter. Therefore each person's parameter estimate is obtained individually and independently of the distribution of person parameters. In applying the software RUMM2030,¹⁴ the ML method, noted earlier, is applied in this paper.

Although the ML equation is not always used in person parameter estimation, for example to account for bias in estimates with close to extreme values weighted likelihood estimates may be used,¹⁵ the same principle as with ML holds. In the RM, the total score $r_n = \sum_{i=1}^I x_{ni}$ is the sufficient statistic for the person estimate β_n . Consequently, irrespective of the pattern of responses, persons with the same total score r_n have the same estimate, notated $\hat{\beta}_r$. Equation 11 is the ML solution equation in which it is evident that the only statistic involving data is the total score r_n :

$$r_n = \sum_{i=1}^I \hat{p}_{ni}, \quad (11)$$

where $p_{ni} = \Pr\{X_{ni}=1\} = \exp(\beta_n - \delta_i) / \gamma_{ni}$ of Equation 1.

The identity of the estimate for the same total score, irrespective of the pattern of responses, holds only for responses to identical items. Otherwise, even with responses to the *same number* of items that are not identical, the same total score will not give the same estimate. As required, for a given total score r , the more difficult items will give an estimate $\hat{\beta}_r$ which is greater than one with easier items. To formalise the estimate with different items, Equation 11 is qualified by Equation 12 in which $h_{ni} = 0,1$ is an indicator variable which takes the value 1 if the person has responded to the item i , otherwise it takes the value 0:

$$r_n = \sum_{i=1}^I h_{ni} p_{ni} \quad (12)$$

3.3 Person estimates at Time 2 in the presence of response dependence from Time 1

Let β_{n1} , β_{n2} be the person locations at Time 1 and Time 2 respectively. Consider again the matrix of responses at Time 2, augmented by the responses $x_{n(k1)}$ to item k at Time 1, with the response $x_{n(k2)}$ resolved with respect to $x_{n(k1)}$. Recall that in this analysis, the responses to item k at both Times 1 and 2, that is $x_{n(k1)}$, $x_{n(k2)}$, which are governed respectively by the person parameter β_{n1} and β_{n2} , are both eliminated. However, because it governs the responses at Time 2, the person location parameter β_{n2} is present in the resultant matrix of the resolved items. Therefore, the proficiencies β_{n2} are the only *person* parameters that govern the responses in the matrix analysed at Time 2, with no role for the proficiencies β_{n1} from Time 1.

Now consider the role of the resolved items $(k2)_{10}$, $(k2)_{11}$ which replace the original item k_{n12} in the estimate of β_{n2} . For illustrative purposes, suppose $d_k > 0$, in which case the response at Time 2 is more likely to be the same as that at Time 1 than it would be if it were governed only by β_{n2} . Thus the responses to the resolved items $(k2)_{10}$, $(k2)_{11}$ are governed, not only by the person parameter β_{n2} at Time 2 and the difficulty δ_k of item k , but also by the response dependence effect d_k which makes similar responses at the two times more likely than they would be without dependence. Thus if person n had the response $x_{n(k1)} = 0$ at Time 1, then the response at Time 2 is to item $(k2)_{10}$, whose difficulty is $\hat{\delta}_{k2,10} = \hat{\delta}_k + \hat{d}_k$. This increase in difficulty by $\hat{d}_k$ of the item k , which increases the probability of a response $x_{n(k2)} = 0$, is exactly the magnitude of the response dependence d_k . Likewise, if person n had the response

$x_{n(k1)} = 1$ at Time 1 then the response at Time 2 is to item $(k2)_{11}$ whose difficulty is $\hat{\delta}_{k2.11} = \hat{\delta}_k - \hat{d}_k$. This reduction in difficulty by $\hat{d}_k$ of item k , which increases the probability of a response $x_{n(k2)} = 1$, is exactly the magnitude of the dependence d_k .

3.4 Person estimates under change and response dependence

To illustrate the implications for the person estimates from the resolution of item k with response dependence between Time 1 and Time 2, suppose that there are a total of five items and that Item 3 has a response dependence of $d_3 = 2$ and is the only item with response dependence. The difficulties of Item 3, not resolved and resolved, are shown in Figure 1. The difficulties of the items are shown in the first set of rows of Table 1. The person estimate $\hat{\beta}_3$ for a total score of 3 for these items at Time 1 is $\hat{\beta}_3 = 0.513$.

Table 1 Person proficiency estimates for a total score $r_2 = 3$: item 3 has a response dependence of $d_3 = 2$ at Time 2.

Original						
Time 1	$\delta_1 (-1.50)$	$\delta_2 (-0.75)$	$\delta_3 (0.00)$	$\delta_4 (0.75)$	$\delta_5 (1.50)$	
$\hat{\beta}_3$	p_{31}	p_{32}	p_{33}	p_{34}	p_{35}	r
0.513	0.882	0.780	0.626	0.441	0.272	3
Resolved 3.1						
Time 2	$\delta_1 (-1.50)$	$\delta_2 (-0.75)$	$\delta_{31} (-2.00)$	$\delta_4 (0.75)$	$\delta_5 (1.50)$	
$\hat{\beta}_{32}$	p_{31}	p_{32}	p_{331}	p_{34}	p_{35}	r
0.143	0.838	0.710	0.895	0.353	0.205	3
Resolved 3.0						
Time 2	$\delta_1 (-1.50)$	$\delta_2 (-0.75)$	$\delta_{30} (2.00)$	$\delta_4 (0.75)$	$\delta_5 (1.50)$	
$\hat{\beta}_{32}$	p_{31}	p_{32}	p_{330}	p_{34}	p_{35}	r
1.018	0.925	0.854	0.272	0.567	0.382	3

The second set of rows of Table 1 shows the difficulty of item 3 in the case that the person has responded $x_{n(31)} = 1$ at Time 1. The difficulty of the item, which originally was 0, is now effectively $\hat{\delta}_{32.11} = 0 - 2 = -2$. This difficulty reflects that the response is not only governed by the person location $\hat{\beta}_{32}$ at Time 2 and the initial difficulty, $\hat{\delta}_3 = 0$, but also by the response dependence effect, which in this case makes the resolved item effectively easier than its original difficulty by exactly the response effect $d_3 = 2$. The

item is made easier in order to increase the likelihood of the same response $x_{n(32)}=1$. As a result, the person's proficiency estimate for a score of $r_2=3$ receives less credit than if there were no effect of dependence. This estimate, $\hat{\beta}_{32}=0.143$, shown in Table 1, is clearly less than $\hat{\beta}_3=0.513$ which is the estimate obtained if only the original item difficulty prevailed at Time 2.

The third set of rows of Table 1 shows the difficulty of item 3 in the case that the person has responded $x_{n(31)}=0$ at Time 1 and $x_{n(31)}=1$ at Time 2. The effect is complementary to that in the second set of rows, with the resolved item difficulty of item 3, $\hat{\delta}_{32.10}=0+2$, more difficult than the original difficulty. This effective increase in difficulty acknowledges in the equation that if a response is 0 at Time 1, then it is more likely to be 0 at Time 2. The item as such, as indicated earlier, has not become more difficult at Time 2. As a consequence, for the same total score of $r_2=3$, the person location estimate $\hat{\beta}_{32}=1.018$ is clearly greater than $\hat{\beta}_3=0.513$. That is, if despite the response dependence enhancing the same response of 0 at Time 2, the response is nevertheless 1, the person receives greater credit than if there were no dependence. In both cases, the change in the difficulty compensates for the effect of dependence. Thus for the testing at Time 2, persons who have the same total score do not necessarily have the same person estimate – the estimate is affected by the response to the same item at Time 1 and is compensated for by exactly the magnitude and direction of the dependence formulated in Equation 5. These effects are illustrated below with sets of simulation studies.

In the above illustrative example with a total of just five items, only one item was considered to have response dependence - the other items had invariant relative item difficulties. However, and conveniently, the principle of resolution shown holds even if all items have dependence and all items at Time 2 are resolved with respect to the responses to the same items at Time 1. The resolution of all items creates a matrix with

missing data for every item, but with a reasonably sized data matrix and with the person proficiencies well aligned to the item difficulties, the missing data is unlikely to produce an ill-conditioned matrix in the sense that it can be divided into mutually exclusive submatrices which would preclude parameter estimation.¹⁶

4. A simulation study

In a series of simulation studies on the measurement of change, Marais¹⁷ showed that in the presence of response dependence in the form of Equation 5 the alignment between the difficulties of the items and the distribution of person proficiencies affected the direction of the bias in the person estimates at Time 2. She also showed that with increasing dependence there was an increase in the likelihood that items would show bias in their difficulty estimates at Time 2. Of course, the biases in the item difficulty and person proficiency estimates are intertwined.

The design of the simulation study of this paper is shown, together with the estimates of the person proficiencies, in Table 2. The number of dichotomous items in all studies was 36 with equal differences between items ranging from -3.5 to 3.5, giving them a mean of 0 and a standard deviation of 2.125. The magnitude of change between means from Time 1 to Time 2 in all simulations, where it is present, is $a = |0.35|$ logits; the magnitude of dependence in all simulations for all items, where it is present, is $d = 1.5$. These orders of magnitude were shown to have observable effects in the Marais study.¹⁷ In addition, in all studies the sample size was a large 5,000. The large sample size ensured estimates of the item parameters had small standard errors. In all simulations the person proficiencies were normally distributed with varying means and with a constant standard deviation of 1.25 logits. Although a large sample was simulated in order to provide accurate item parameter estimates, because all estimates have associated standard errors and because this is not a study that assesses the distribution of

estimates of the person parameters under different conditions, replications with exactly the same parameters were not carried out.

The parameter d_i is a relatively new parameter in the literature. Therefore, because its estimate for each item i appears naturally with the different items, and it estimates the same value for all items, the mean and standard deviation of the replicated estimates in each simulation is reported and compared to the simulated value.

To provide a frame of reference, the first simulation set of three studies had 0 change, $a = 0$, and 0 dependence, $d = 0$, with differences in the initial means of the persons among the three studies. A second simulation set of three studies had a change of $a = 0.35$ logits from different initial means, and again with 0 dependence, $d = 0$. These two studies demonstrate that resolving the matrix at Time 2, in the presence of no response dependence, produces no bias in the proficiency estimates. This confirms that any effects in the presence of response dependence can be attributed to response dependence and not to the method of analysis.

The third and fourth simulation sets of three studies each had a dependence of $d = 1.5$ for all items. Because the resolution of responses at Time 2 is carried out for each individual item, there is no need to have the same dependence in all items. However, for the purpose of this illustration, and to confirm the accuracy of the estimates of the dependence value and that of the change, the dependence was made equal among items. Thus in the third simulation set of three studies there was 0 change, $a = 0$, but with different means at Time 1 and dependence $d = 1.5$. In the fourth simulation set of three studies there was a change of $a = |0.35|$ and again a dependence of $d = 1.5$. The critical check on the quality of estimates is that shown in the last set of studies, where there is both change and response dependence.

The analyses of all data sets involved the same, following steps. First, the data at Times 1 and 2 were analysed separately, and the estimated means and standard deviations compared. They were also compared to their known simulated values. Second, the responses at Time 2 were resolved, and again the means and standard deviations compared to the known theoretical values. In addition, as a bi-product, the estimate of the response dependence $d = 1.5$ was calculated for each study. Finally, an estimate of change, in terms of the difference between means at Times 1 and 2, was calculated. For completeness, a t test for significance of difference between them was formed. With a sample size of 5,000, in the presence of change, the dependent sample t test was expected to be significant. However, its main relevance is in the examples where there is no change but in which there is response dependence. In these cases, a non-significant change and an accurate estimate of the response dependence contributes to the validation of the logic of the procedure for controlling response dependence. Otherwise, even with no change, response dependence can give the impression of change. In addition, the estimates of the means and standard deviations at Time 2 in all cases are compared to their known, simulated values.

5. Results

Table 2 shows the results from the four simulation sets, each with three studies. Because the simulated means will not have exactly the same values as the theoretical distributions from which they are simulated, they are shown for both the simulations at Times 1 and 2. The responses at Time 1 and Time 2, given the respective person parameters and generated from the normal distribution with specified mean and standard deviation, are simulated independently. In the case of change for Time 2, each person's parameter is increased by $a = |0.35|$.

Table 2 Design of the simulation studies with the results of the estimates of the person proficiencies.

INSERT TABLE 2

5.1 Simulation Set 1 - no change and no dependence

Simulation Set 1 has no change, $a = 0$, and no dependence, $d = 0$. The means for Studies 1, 2 and 3 at Time 1 are -0.35 , 0.00 , 0.35 respectively. It is evident from Table 2 that the simulated values, -0.348 , 0.002 , 0.352 , for both times (which in the case of no change are identical), are close to their theoretical values. These are the values to which the estimates are compared. Without repeating them, the same closeness to their theoretical values is evident for the simulated standard deviations. The estimates of the means at both times, and in all three studies

$(-0.347, -0.348; 0.005, 0.000; 0.347, 0.351)$ are also close to their simulated values.

This of course is as expected. However, important for setting the frame of reference for the analysis of resolved items at Time 2, is that the same means are effectively estimated from the resolved items at Time 2 $(-0.351, 0.002, 0.352)$. Therefore the estimate of the change, both with the unresolved and resolved items in the case of no dependence and no change between Times 1 and 2, will be effectively 0. This is shown by the t test values at the right of Table 2, none of which are significant, even with the sample of size of 5,000. Thus resolving items with no response dependence does not distort the estimates.

The standard deviations of the estimates are slightly larger than those of the simulated values which is expected because the estimates contain error. A correction to the variance is possible from the standard errors of the estimates that are available from the ML estimate for each person. The mean error variance $\bar{\sigma}_e^2$ of the ML estimates is given

by $\bar{\sigma}_e^2 = (\sum_{n=1}^N \sigma_{\beta n}^2) / N$ where $\sigma_{\beta n}^2$ is the error variance for each person n 's estimate.

Then the corrected estimate of the variance, notated in Table 2 as $\hat{\sigma}_\beta^2$, is given by

$\hat{\sigma}_\beta^2 = \hat{\sigma}_{\hat{\beta}}^2 - \bar{\sigma}_e^2$. Table 2 shows the square root of this value, $\hat{\sigma}_\beta$. It is clear that for all

three studies in Simulation Set 1, these corrected estimates,

(1.247, 1.235; 1.245, 1.236; 1.240, 1.241) for Times 1 and 2 are close to the simulated

values and close to the theoretical value of 1.25. They are also close in the resolved

analysis (1.238; 1.237; 1.238). These values are relevant to report because in the

presence of dependence, not only the means, but the standard deviations at Time 2 are affected.¹⁷ This effect is shown in the simulations with dependence.

Table 2 also shows the estimates of the dependence effect d from Equation 5. As

required, the estimate is close to 0, confirming the excellence of its estimate

demonstrated by Andrich and Kreiner.¹³ Two standard deviations for the estimate are

shown: $\sigma_d^{(1)}$ is the standard deviation of the mean of the estimates from all items; $\sigma_d^{(2)}$

is calculated from the standard errors of the estimates shown in Equation 10. It is clear

that they are very close to each other, confirming the stability of all estimates concerned with response dependence.

5.2 Simulation Set 2 - change but no dependence.

Simulation Set 2 has a change of $a = 0.35$ from three mean locations at Time 1

(-0.35, -0.175, 0.00) for Studies 1, 2, and 3. In these studies, there is no dependence,

$d = 0$. Again the estimates of the means and standard deviations are as expected, and

therefore immediate focus is on the estimate of the change of 0.35 in each of the three

studies at Time 2. It is evident that both the unresolved and the resolved estimates at

Time 2, (-0.033, -0.027; 0.141, 0.142; 0.341, 0.317) for the three studies, are not only

close to each other, but also close to the simulated values of $(-0.024, 0.151, 0.326)$. As a result, the t test values, of the order of 30, show highly significant change and the values for the unresolved and resolved change are virtually identical. Again, as required because no dependence was simulated, the estimates of d were virtually 0 and not significantly different from 0.

The studies in Simulation Sets 1 and 2 confirm that in the presence of no dependence, the proposed procedure of resolving items at Time 2 depending on the response at Time 1 has no effect on the estimates of change. As commented above, this confirmation assures that any differences obtained with dependence is not a result of problems with the estimation procedure with resolved items as such, but results from the presence of dependence itself.

5.3 Simulation Set 3 – no change but dependence

Simulation Set 3 has no change in any of the studies, which have the three different theoretical means at Time 1 $(-0.35, 0.00, 0.35)$ and a dependence of $d = 1.5$. In these studies some evidence of dependence in the estimates is expected. The simulated and estimated values at Time 1 and the simulated values at Time 2 are as expected.

However, the estimates at Time 2 $(-0.374, -0.024, 0.326)$ in the three studies compared to their simulated values $(-0.477, 0.019, 0.427)$ show noticeable bias in the first and third studies. Indeed, for the first and third studies, the t values are significant even though no change was simulated, while for the second study, the t test is not significant. Two observations follow from the above results. First, that where the mean person location is well aligned to the items, in which case the responses at Time 1 have a more or less symmetrical range of scores around the mean of the location of the items, then the effect of dependence for the different persons seems to cancel out each other and the mean remains stable. On the other hand, where the mean of the persons is below the mean of the items, the response dependence provides an estimate which suggests a negative change even though there is no change; symmetrically, where the mean of the persons is above the mean of the items, the response dependence provides an estimate which suggests a positive change when there is no change. These effects are consistent

with those found by Marais.¹⁷ In addition, the estimated standard deviations, corrected for error, have the values (1.610, 1.622, 1.611), which are also noticeably greater than their simulated values, 1.256 in each case. Even in the case where the person locations are aligned with the item locations and there is no impact on the means, the standard deviations are greater than their known values.

In contrast, the estimates from the resolved data (–0.368, –0.021, 0.326) are all close to their simulated values (–0.374, –0.024, 0.326), including those of the first and third study in the set. Indeed as required from a correct analysis, the t tests show no significant change, consistent with simulation of a change of 0. In addition, the standard deviations (1.247, 1.268, 1.255) are very close to the simulated value of 1.256. Finally from this set, the estimated values of $\hat{d}$ (1.480, 1.484, 1.491) are very close to 1.50. The two estimates of the standard deviations of $\hat{d}$ are also very close to each other, having a maximum difference of only 0.003. The analysis of the resolved matrix confirms the removal of the systematic bias in the person estimates that arises when the mean location of the persons deviates from a close alignment to the mean of the item difficulties. This removal of bias is consistent with the mathematics and the logic of the resolution of items at Time 2 given responses at Time 1.

5.3 Simulation Set 4 – change and dependence

Simulation Set 4 of three studies has a change of $a = 0.35$ and a dependence of $d = 1.5$. In studies 1 and 2 the change is from respective simulated means (–0.359, –0.184), and in study 3 the change is $a = -0.35$ from a positive simulated mean of 0.341. Because they contain both change and response dependence, the results of these studies are considered somewhat more closely than those of the previous sets.

In study 1, where the simulated mean at Time 1 is -0.350 and is below the mean of the item locations of 0.0 , the simulated value at Time 2 is -0.009 , and the unresolved estimate is -0.199 . The latter estimate shows that the effect of dependence in this case is to have a mean closer to the mean at Time 1 than that simulated. In contrast, again the estimate from the resolved matrix, -0.001 , is very close to the simulated value. In study 2, where the simulated mean at Time 1 is -0.184 and that at Time 2 is 0.166 , both means are relatively well aligned to the mean of the item locations and the unresolved and resolved estimates are respectively $(0.018, 0.167)$. These again show a regression towards the mean at Time 1 for the unresolved estimate and extremely accurate resolved estimates. In study 3, the mean at Time 1 is 0.341 and is above the mean of the item locations, and the simulated value at Time 2 is -0.009 . The unresolved estimate is 0.204 which again shows the effect of dependence, producing an estimate closer to that at Time 1 than that simulated. In contrast, again the estimate from the resolved matrix, 0.007 , is very close to the simulated value. The t values, from the unresolved and resolved matrices, show significant t tests, but clearly the estimates from the resolved analysis are very close to those of the simulated values.

Because it is the most telling case in demonstrating the control of response dependence, with the presence of both change and dependence, the estimates of the change when compared to the simulated values are shown below. The simulated change for the first two studies are both 0.350 , that is $(-0.009 - (-0.359))$ and $(0.166 - (-0.184))$, and for the third study the change is -0.350 , that is $(-0.009 - 0.341)$. The estimates of change from the unresolved matrix in the first two studies are respectively $0.153, 0.197$ which are noticeably smaller than 0.350 . In the third study the estimate of change is -0.143 , which is also noticeably smaller in magnitude than -0.350 . In each case the change estimated is smaller than was simulated, showing, not surprisingly, a regression to the mean at Time 1. The regression arises from the response dependence in which the response at Time 2 tends to be more like the response at Time 1 than it would be if it

were governed only by the person parameter at Time 2. On the other hand, the estimates from the resolved matrix for the first two studies, (0.351, 0.347), and in the third study (–0.341), show no systematic regression. The evidence from Simulation Set 4, that response dependence is controlled and eliminated in the estimate of change through the resolution of the matrix at Time 2 relative to responses at Time 1, is compelling.

Where there is positive change overall, the elimination of the effect of dependence occurs as illustrated in the third set of rows of Table 1. That is, those persons whose response changes from 0 to 1 receive more credit than they would otherwise, and in the presence of positive change, and despite response dependence, there are more such people than those whose response is the same.

The standard deviations are again accurate in the resolved analysis while noticeably larger than they should be in the unresolved analysis – in the former they are respectively 1.260, 1.266, 1.267 relative to a simulated value in each case of 1.248, while in the latter they are larger, (1.613, 1.615, 1.620). Finally, in the presence of change, the estimate $\hat{d}$ (1.468, 1.479, 1.483) is very accurate and as accurate as when there is no change.

6 An example with real data

The example with real data is from the Functional Independence Measure (FIM), a clinician-scored measure of functioning, introduced by Keith, et al.¹⁸ Lundgren, Nilsson and Tennant¹⁹ reviewed its properties using RM analyses. The data from the FIM used in this paper contains 13 items concerned with degrees of physical functioning. Each item has seven putatively ordered categories with the greater scores indicating better proficiency in physical functioning. Data were extracted from a

database of outcome measures recorded prospectively during patients' stays at an inpatient rehabilitation unit.²⁰ Data had originally been collected as part of a long-term data collection project. The FIM was scored by members of the multidisciplinary rehabilitation team within three days of admission to the unit and in the two days prior to the patients' discharge from the unit. The complete data set includes patients in four diagnostic groups: stroke, spinal cord injury, multiple sclerosis (MS), and other neurological conditions.

6.1 Adaptations of the data

For the illustrative purposes of this paper, the following adaptations to the data were made. First, only the 602 patients with MS and complete data at both times of assessment were considered. In this group there were 402 females and 200 males. Using the MS only patients made the group more homogeneous with respect to the responses and therefore reduced potential group effects that might have interacted with response dependence and change. Second, the items were dichotomised with responses in the first five categories combined to be given a score of 0, and the remaining two categories combined and given a score of 1. The particular dichotomisation had the persons, with a mean of 0.113, well aligned to the mean item difficulty of 0.0. Although dichotomisation typically requires detailed justification, in this case it is justified in part because the analysis is used illustratively for dichotomous responses, and in part because there is evidence that there are too many categories for all to be working consistently. Third, because it had an extremely poor discrimination at both times, item 12 (concerned with walking) was eliminated at both times from the data matrix. The remaining items at Time 1 were entirely compatible with the RM.

6.2 Item difficulty and dependence estimates from the real data

In the simulated data, the main focus was on the estimation of the person estimates under conditions of change and response dependence and referenced to known values of the distributions. The estimate of dependence in each simulation, which was common for all items for each simulation, was also shown. In a real data set, it is instructive to report the item difficulty parameter estimates, which include any effects of dependence, before considering the effects of dependence on person proficiency estimates.

Table 3 shows the item difficulty estimates for Time 1 data, Time 2 data analysed independently of the Time 1 data, and from the analysis in which the Time 2 responses for each item were resolved with respect to responses to the same item at Time 1. The last of these was calculated from Equation 10a by analogy to the calculation of the estimate for dependence. Thus instead of taking the difference between the two estimates in Equation 10a to obtain an estimate of d_k , the sum of the two estimates gives the difficulty estimate δ_k :

$$\delta_k = (\hat{\delta}_{k2.10} + \hat{\delta}_{k2.11}) / 2, \quad (13)$$

with a standard error

$$\sigma_{\hat{\delta}_k} = \sqrt{(\hat{\sigma}_{\delta k2.10}^2 + \hat{\sigma}_{\delta k2.11}^2) / 4}. \quad (14)$$

Finally, Table 3 also shows the estimated dependence values of d_k for each item.

Table 3 Estimates of three sets of difficulties δ and of the dependence d for each item

Item	T1	SE1	T2	SE2	Res*	SE Res	d	SE d
1	-0.437	0.117	-1.325	0.154	-0.971	0.175	0.785	0.175
2	-1.887	0.128	-2.048	0.166	-1.634	0.214	1.193	0.214
3	0.277	0.117	0.578	0.135	0.839	0.149	0.820	0.149
4	-1.548	0.124	-1.555	0.157	-1.167	0.196	1.216	0.196
5	0.708	0.118	0.919	0.132	0.464	0.226	1.748	0.226
6	-0.855	0.119	-0.325	0.144	-0.303	0.214	1.416	0.214
7	0.169	0.117	0.410	0.137	0.681	0.141	0.927	0.141
8	-1.695	0.126	-1.561	0.157	-0.817	0.167	1.195	0.167
9	0.437	0.117	0.183	0.139	-0.036	0.223	1.366	0.223
10	-0.027	0.117	-0.466	0.145	-0.878	0.300	1.622	0.300
11	1.878	0.126	1.995	0.127	1.675	0.162	1.333	0.162
13	2.979	0.150	3.194	0.139	2.147	0.231	1.972	0.231
$\bar{\delta}$	0.000		0.000		0.000		1.299	
σ	1.445		1.558		1.794		0.362	

Res*: Resolved analysis

Interpreting first the estimate of dependence, it is evident that every item in Table 3 has statistically significant response dependence with the ratio $d / SE d$ substantially greater than 2 (significance at the level of 0.01) for each item. The average value of d for the items is 1.299 with a standard deviation of 0.362 indicating some variation among the items in the degree of response dependence. However, it is noteworthy that all items have response dependence. Why this is the case cannot be answered from the statistical analysis itself. Some factors regarding the patients, clinicians, or an interaction between

the two, leads to a response that is more similar at Time 2 to Time 1 than it would be if only the parameters of the patients at each time and of the items were governing the responses at both times. Second, the standard deviations of difficulties at Time 1 and 2 are similar, with that of the resolved estimates being greater. The greater similarity between the first two estimates arises in part because of the response dependence.

INSERT FIGURE 2

Figure 2. Unresolved and resolved item difficulty estimates at Time 2 relative to those at Time 1.

6.3 *Considerations of differential item functioning*

Because differential item functioning (DIF) between times can impact real data, a summary discussion of its issues are considered before showing the final proficiency estimates when account is taken of response dependence. The presence of DIF between two times of assessment implies that the relative difficulties of the items at the two times are not the same. Reference is to *relative*, rather than *absolute* difficulties, because in each analysis, the sum of the difficulties of the items is 0.0. The relative item difficulties shown in Table 3 suggest there is some DIF between the two times and between the items resolved for dependence and Time 1. To summarise this DIF, Figure 2 shows plots of the unresolved and the resolved estimates at Time 2, both relative to Time 1. The line of unit slope provides a frame of reference for the hypothesis of no DIF. It is evident that there is greater deviation from the line of unit slope, and therefore greater DIF, with the estimates of difficulties of the resolved items than of the unresolved ones. The main aspect of DIF is the regression of difficulties towards the arbitrary mean of 0.0. However, as with the standard deviations, the reduced DIF at Time 2 arises in part because of the response dependence between the two times.

The DIF detected by the relative differences in means shown above is only of the difficulty and is not as elaborate as when DIF is considered across the continuum. The role of DIF and how it is to be dealt with depends on the context. For example, in linking designs it is essential that the items used for links show no DIF.^{21,22} In the example of this paper, the DIF from Time 1 to Time 2 arises simply because the average change in the functioning of the group of persons is not uniform across the items. This may arise, not because of some fault in the instrument or the rehabilitation, but because some aspects of functioning may have been deemed more important than others at the time, and greater emphasis was placed on the rehabilitation in those areas. Of course some other aspects that received relatively less attention might have needed relatively more. It may also be that some aspects are easier on which to improve. The analysis discloses these features but information external to the analysis is required to decide upon any action because of the DIF. In some cases it may be more valid for the comparisons to ignore DIF than to take account of it.²³

However, DIF is relative, and because there is a constraint on the estimates of the item difficulties in each analysis, if some item does have real DIF because of the emphasis on the rehabilitation of the aspect assessed by that item, and therefore appears to be relatively easier, then even if there is no real DIF for the other items, they will show relatively greater difficulty. This effect gives rise to artificial DIF, explained in some detail by Andrich and Haquist.²¹ In addition, although the fit of the data to the model will be better than if the item difficulty parameters from Time 1 are used at Time 2, the item parameters between the two assessments are no longer invariant. Nevertheless, because both analyses have the same constraint of 0 on the item difficulties, in some cases and for some purposes this may be a legitimate analysis.

In other cases it might be considered that the most valid assessment of change is to anchor the relative difficulties of items at Time 1 in estimating the proficiencies of

persons at Time 2. This ensures invariance of item difficulties between the two times. However, because the parameters from the second time of testing, consistent with the data, are not used, the fit will be worse than if the parameter estimates from Time 2 were used. Of course the degree to which it is worse depends on the degree of DIF between the times – if there were no DIF, then the fit would not be affected. Which method is more valid in the presence of DIF must be determined outside the statistical analysis itself. In some cases the two methods might give very similar results especially if the same items are used in the different analyses, in others they may not. In addition, the lack of uniform change for any person can be diagnosed from the person-fit indices. There may be ready explanations why a person's profile at Time 2 is not summarised by a single proficiency estimate.

The point of noting the above possibilities is to recognise that DIF is a relatively complicated matter and that even in the illustrative analysis of this paper, which is not considered definitive for the data, DIF cannot be ignored entirely.

6.4 *Measurement of change in proficiency in the real data*

For the purpose of this paper, results from the analyses of person proficiencies at Time 2 with four different item parameter estimates are presented. In these analyses, the focus is on the difference when the resolved analysis is used compared to the other possible analyses.

Table 4 shows the person estimates from the following five analyses. The first is from the analysis of Time 1 data. Change is referenced to the mean estimate from this analysis. The remaining four analyses are from the Time 2 responses from (i) estimates of item difficulties from Time 2, (ii) estimates of item difficulties anchored to those from Time 1, (iii) item difficulties at Time 2 obtained from items resolved for

dependence relative to Time 1, and (iv) item difficulties at Time 2 obtained from items resolved for dependence relative to Time 1 but with the mean difficulties of each pair of resolved items anchored to the value at Time 1. In each analysis, the same constraint, $\sum_{i=1}^I \hat{\delta}_i = 0$, is present. The Table also shows the difference in mean estimate from the four analyses at Time 2 and the mean estimate at Time 1.

The method by which the item difficulties can be anchored to those from Time 1 for the last analysis, while simultaneously accounting for response dependence at Time 2, is to apply Equation 10a. That is, the difficulty $\hat{\delta}_k$ estimate from the Time 1 analysis is entered into Equation 10a, with the estimates of $\hat{d}_k$ given by the resolved analysis according to Equation 10b. Because the resolved items have the same magnitude of deviations but opposite signs from the common difficulty giving $\hat{\delta}_k \pm \hat{d}_k$, the mean difficulty of the items and the origin of the scale remains $\sum_{i=1}^I \hat{\delta}_i = 0$. Because the item difficulties are anchored to those in the Time 1 analysis, it indicates that both the origin and unit are the same in the two analyses. This last analysis may be considered the most valid in terms of scale in that the effect of dependence has been eliminated while retaining the same origin and unit of the analysis as Time 1.

Table 4 Person estimates from Time 1 and from four item scales at Time 2

	Time 1	Time 2	Time 2 anchored to Time 1	Time 2 resolved for dependence	Time 2 resolved for dependence and anchored to Time 1
Mean	0.113	1.352	1.291	1.665	1.777

(Time 2 – Time 1)		(1.239)	(1.178)	(1.552)	(1.664)
St. Deviation	2.266	2.541	2.446	2.222	2.516

Whether anchored or not to the Time 1 difficulties, the estimates for Time 2 are similar (1.352,1.291) when items are not resolved. However, in both resolved cases the estimates (1.665,1.777) have a greater mean. Considered first is possibly the standard comparison - the difference between means from the Times 1 and 2 analyses with the Time 2 data anchored to the Time 1 difficulties. Clearly there is a substantial improvement ($1.291 - 0.113 = 1.178$) logits, which is clearly statistically significant (dependent $t = 21.073$).

Considered second are the novel comparisons – the difference between means from Times 1 and 2 with the Time 2 data resolved to take account of dependence between Times 1 and 2. First, the simple resolution without anchoring on the Time 1 estimates has a difference between Times 1 and 2 of ($1.665 - 0.113 = 1.552$) logits which is greater than the difference from the previous estimate (1.178). Second, with item difficulties resolved and their mean anchored to those at Time 1, the difference of 1.664 ($1.777 - 0.113$) is even greater than 1.552. This is as expected from evidence of response dependence in the data, the logic of accounting for it and evidence from the simulation study which shows that the difference between times of assessment is likely to be reduced in the presence of response dependence. From the most comprehensive analysis, the fourth, where dependence is taken into account and the mean difficulties are anchored to Time 1 and when no account is taken of response dependence, the difference is virtually 0.5 logits ($1.777 - 1.291 = 0.486$). In this example, the difference between the analysis at Time 2 when the Time 1 difficulties are anchored only and when response dependence is also accounted for is statistically significant ($t = 11.315$). Thus in this illustrative example, the presence of response dependence substantially underestimates the magnitude of improvement between Times 1 and 2 of assessment.

It is stressed that the analysis of the real data is illustrative and that it is not a definitive study of the change based on the FIM for this data set or in general. Different data sets, even with the same instrument, may show different degrees of dependence.

Nevertheless, the paper shows a method by which response dependence between time points of assessment on the same items can be controlled, quantified, and its effect on change estimated using the RM.

In the introduction to controlling response dependence in this paper, the focus was on dichotomous items. In many areas of assessment, items have more than two ordered categories. The method of resolving items and the estimation of the response dependence for polytomous items has been generalized by Andrich, Humphry and Marais.²⁴ The generalization to assessing and removing response dependence over two times of assessment with the same items is readily generalized. This generalization is left for other occasions.

Comment

The concern with local response dependence is brought out explicitly by the application of the Rasch model and the requirement of statistical independence. However, response dependence is a property of the data, and if the method of analysis for comparing estimates at different times does not control or eliminate the response dependence, it affects all analyses and subsequent interpretations. As demonstrated in this paper, response dependence between two times of assessment on the same items will generally attenuate the magnitude of the change, with the degree of attenuation affected by the alignment of the person estimates relative to the item difficulties. The presence of response dependence in an example of the analysis of a real data set, modified for illustrative purposes, suggests that in the assessment of change over time when the same items are used at the multiple times of assessment, the effect of response dependence

should be assessed. The possible difference, either the effect of a treatment or in monitoring improvement, may be substantially underestimated with important implications for individuals, treatments, and policy decisions. Where the facilities are implemented into the software for the resolution of the responses, as they are in RUMM2030,¹⁴ then the assessment can be made relatively readily by simply constructing a person-by-response matrix in which the responses at the two times provide two sets of item responses for each person. An added benefit of the procedure is that it is not necessary to identify items that may or may not have dependence – this is identified in the process of assessing response dependence.

Acknowledgements

Discussions with Jeremy Hobart, Stefan Cano and Alan Tennant regarding the importance of assessing and controlling response dependence in repeated assessments in health outcomes motivated the first part of the research reported in this paper. The work was enhanced at a workshop over two days in the Department of Rehabilitation studies at The University of Leeds in which Maren Boecker, Curt Hagquist, Liz Hensor, Mike Horton, Rory O'Connor, Thomas Salzberger, Anita Slade, Alan Tennant, and Mark Wilberforce participated and in which the analysis of the real data set reported in this paper was studied. Rory O'Connor kindly provided the data for the real example. Ida Marais, Irene Styles and Sonia Sappl read the paper and made it more efficient, and Sonia Sappl carried out the simulations. Two anonymous reviewers provided constructive suggestions.

The data of the simulation studies and the real data analysed for this paper are available at <https://researchdataonline.research.uwa.edu.au/handle/123456789/3221>

Andrich, D., (2016): Controlling Response Dependence in the Measurement of Change using the Rasch model. The University of Western Australia.

<http://dx.doi.org/10.4225/23/583268a4e4283>

Declaration of conflicting Interests

The Author declares that there is no conflict of interest.

Funding

This work was supported by the Australian Research Council [linkage grant number LP140100567] with the School Curriculum and Standards Authority of Western Australia and the Australian Curriculum, Assessment and Reporting Authority as Industry Partners; and Pearson with a Psychometric Research Endowment.

References

1. Harris CW, editor. *Problems in measuring change*. Madison: University of Wisconsin Press; 1963.
2. Rasch G. *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press; 1960.
3. Wright BD. A history of social science measurement. *Educational Measurement: Issues and Practice* 1997; 16(4): 33-45.
4. Embretson SE. Item response theory models and inferential bias in multiple group comparisons. *Applied Psychological Measurement* 1996; 20: 201-212.
5. Fischer GH. An IRT-based model for dichotomous longitudinal data. *Psychometrika* 1989; 54(4): 599–624.
6. Andersen EB. Conditional inference for multiple choice questionnaires. *Br J Math Stat Psychol* 1973; 26: 31-44.
7. Zwinderman AH. Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement* 1995; 19(4): 369-375.
8. Gulliksen H. *Theory of Mental Tests*. New York: Wiley; 1950.
9. Lazarsfeld PF and Henry NW. *Readings in Mathematical Social Science*. Chicago: Science Research Associates Inc; 1966.
10. Andrich D. Essay review of R. Langeheine and J. Rost, *Latent Trait and Latent Class Analysis*, New York, 1988. Plenum Press. *Psychometrika* 1991; 56: 155-168.
11. Marais I and Andrich D. Formalising dimension and response violations of local independence in the unidimensional Rasch model. *J Appl Meas* 2008; 9(3): 200-215.
12. Kuncel RB and Fiske DW. Stability of response process and response. *Educational and Psychological Measurement* 1974; 34: 743-755.
13. Andrich D and Kreiner S. Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement* 2010; 34: 181-192.

14. Andrich D, Sheridan BS and Luo G. RUMM2030: An MS Windows computer program for the analysis of data according to Rasch Unidimensional Models for Measurement. Perth, Western Australia: RUMM Laboratory; 2015.
15. Warm TA. Weighted likelihood estimation of ability in item response theory. *Psychometrika* 1989; 54: 427-450.
16. Fischer GH. On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika* 1981; 46(1): 59-77.
17. Marais I. Response dependence and the measurement of change. *J Appl Meas* 2009; 19(1): 17-29.
18. Keith RA, Granger CV, Hamilton BB and Sherwin FS. The functional independence measure: a new tool for rehabilitation. *Adv Clin Rehabil* 1987; 1: 6-18.
19. Lundgren Nilsson Å and Tennant A. Past and present issues in Rasch analysis: The Functional Independence Measure (FIM™) revisited. *J Rehabil Med* 2011; 43: 884-891.
20. O'Connor RJ, Beden R, Pilling A and Chamberlain M. What reductions in dependency costs result from treatment in an inpatient neurological rehabilitation unit for people with stroke? *Clin Med (Lond)* 2011; 11(1): 40-43.
21. Andrich D and Hagquist C. Real and Artificial Differential Item Functioning. *Journal of Educational and Behavioral Statistics* 2012; 37(9): 387-416.
22. Looveer J and Mulligan J. The efficacy of link items in the construction of a numeracy achievement scale - from kindergarten to year 6. *J Appl Meas* 2009; 10(3): 247-265.
23. Andrich D and Hagquist C. Real and Artificial Differential Item Functioning in Polytomous Items. *Educational and Psychological Measurement* 2015; 75(2): 185-207.
24. Andrich D, Humphry S and Marais I. Quantifying Local, Response Dependence Between Two Polytomous Items Using the Rasch Model. *Applied Psychological Measurement* 2012; 36: 309-324.