Instability of Person Misfit and Ability Estimates Subject to Assessment Modality*

Alexandra Petridou and Julian Williams
University of Manchester

Abstract

Unexpected response patterns on tests and their problematic interpretation has led to an intense research activity (i) to investigate the sources of such responses and (ii) to model statistics that claimed to detect person misfit in an examinee's response pattern. The rationale behind this effort was the claim that the test scores of these examinees with unexpected response patterns may fail to provide a useful and valid measure of their ability. In this study we have followed-up 'misfitting' examinees in a Mathematics test and during interviews we have asked them to work out items again on which they have provided unexpected responses. Pupils' response patterns were then changed based on their responses during interviews and fit analyses were re-run in order to obtain new estimates of their abilities and of the fit statistics. When old and new estimates were compared using a paired t-test, there was an overall increase in ability estimates and an overall reduction in pupils' Infit and Outfit values. However only the change in the Infit values was statistically significant. By examining the change in ability estimates for each individual pupil we have found cases where the two measures were seriously discordant, raising questions about the validity of test-measurement for these pupils.

Introduction

When examinees take a test their responses are expected to conform to some standard of reasonableness (Smith, 1986): for instance, wrong answers to `easy' questions but right answers to `hard' questions would be regarded as `aberrant'. Such aberrant response patterns

* The authors are grateful for the financial support of the ESRC (PTA-030-2004-00072), the A.G. Leventis Foundation and the University of Manchester.

would be signalled by a high `misfit' statistic computed from the deviations of these responses from those `expected'.

The objective of person-fit measurement is to detect individuals which their item-response patterns are improbable given an IRT model. Person-fit is a question about the meaning i.e. the validity of the measure for specific individuals. According to Smith (1987) no matter how hard we try to construct potentially valid tests there will always be individual performances for whom those tests were not valid. This is because test responses are a function of not only of the items, tasks or stimulus conditions but of the persons responding and the context of measurement. Attempts to find methods of systematically identifying individuals that did not perform as expected led to the development of person-fit statistics. The rationale behind this effort was the claim that 'misfitting' individuals were 'mismeasured' and therefore the scores of these examinees may fail to provide a valid measure of their performance. These claims however were taken for granted by the researchers who focused their efforts on the generation of new indices for the identification of aberrance. No research that proves these claims is present in the literature today. Meijer and Sijtsma (2001) argued that "simply because a response string is improbable does not mean that it is misleading and vice-versa" (p.823).

This paper is part of a larger study which aimed to identify the reasons that lead examinees to 'misfitting' response patterns. Specifically in previous work (Petridou and Williams, under review) we have examined the effect of a number of background variables (e.g. gender, language, anxiety, motivation and ability) on person fit using real data under the framework of a two level (person and classroom) model. For the purposes of this study two person-fit statistics were used i.e. Infit and Outfit MNSQ in order to examine whether the pupils' data were consistent with the Rasch model. The Infit and Outfit values then became the response variables in a one-level and then to two-level logistic. We found that the proportion of misfit attributable to the class level was high at least for the Infit model, ability had a statistically significant contribution to misfit and finally language had also a statistically significant effect but only in the Infit model. A more qualitative work (Petridou and Williams, 2006) followed this study where interviews of 'misfitting' pupils and their teachers took place in order to elicit insights into the causes of their statistical 'aberrance'. Specifically in this study (Petridou

and Williams, 2006) we were looking for 'grounded' explanations of unexpected responses from pupils and their teachers.

During interviews 'misfitting' pupils were asked not only to comment/explain unexpected responses in the test but also to work out again items on which they have provided statistically unexpected responses. Interview data raised concerns in relation to the ability estimates obtained in the test based on what pupils showed to be able to do in the interview situation. Table 1 provides evidence for these concerns as it shows clearly that pupils during interviews were able to work out the majority (i.e. 62%) of items on which statistically unexpected responses were obtained in the test.

Table 1: Cross-tabulation of test versus interview responses

	Test Responses		
Interview outcome	Statistically unexpectedly *correct* responses	Statistically unexpectedly *wrong* responses	**Total**
Statistically unexpected responses that pupils *could not* work out correctly.	67	19	**86**
Statistically unexpected responses that pupils *could* work out correctly	63	80	**143**
Total	**130**	**99**	**229**

Table 1 also shows that pupils were able to work out the vast majority of statistically unexpectedly wrong responses (i.e. 81%) in the interview situation while they were able to work out approximately half of the statistically unexpectedly correct responses.

This paper aimed to examine the stability of ability estimates as obtained in the test subject to change in assessment modality.

Methodology

In order to examine whether the pupils interviewed were misrepresented by the test based on their interview responses, pupils' response patterns in the test were changed according to

pupils' responses in the interview situation. The purpose was to examine whether these changes would result in significant changes in ability estimates and fit statistics. Specifically in cases of statistically unexpectedly wrong responses if pupils during the interviews showed they were able to work out the specific items their test responses were changed from wrong (i.e. code 0) to correct (i.e. code 1) while if pupils were unable during the interview to work out the specific items their responses were left unchanged. The same procedure was followed for statistically unexpectedly correct responses i.e. if pupils during the interviews showed to be able to work out the specific items their test responses were left unchanged while if they were unable during the interview to work out the specific items their responses were changed from correct (i.e. code 1) to wrong (i.e. code 0).

Fit analyses were re-run in order to obtain the new estimates of the abilities and of the fit statistics of pupils interviewed. For the purposes of fit-analyses two general-purpose fit statistics were used i.e. Infit Mean Square and Outfit Mean Square. The new ability estimation was anchored on the previous ability estimates of the rest of the sample, to ensure that the two fit-analyses would produce directly comparable results. Old and new estimates were compared using a paired t-test.

Data and its context

The test data came from a Year 5 mathematics test, developed by the Mathematics for Learning and Teaching (MALT) project of the University of Manchester, which collects diagnostic information and standardize mathematics tests for years Reception to nine. The test was designed to cover the full range of levels and content of the mathematics programme of study for Year 5 (aged 10-11). The initial dataset size was 674 pupils nested within 36 classes coming from 23 primary schools in England. The number of pupils interviewed was 31 nested within 20 classes within 13 schools. Interviews took place approximately 2-3 months after the administration of the test.

Results

Table 2 presents both the old and the new estimates obtained for each pupil interviewed. Table 2 shows clearly that the Infit values were reduced a great deal while the ability estimates for some were decreased while for others were increased. Paired t-tests were

employed in order to examine whether any differences in ability estimates and in fit statistics values were statistically significant.

Table 2: Ability estimates and fit-statistic values before and after changes in response patterns based on what pupils were able to do during the interviews

Pupils	Estimated Ability		Infit values		Outfit values	
	Before	*After*	*Before*	*After*	*Before*	*After*
A2	-3.8	-4.37	1.18	1.04	3.95	7.36
T*	-3.34	-	1.31	-	2.26	-
M6	-3.34	-4.37	0.94	0.88	3.51	0.27
S3	-2.99	-3.65	1.23	1.12	2.11	3.81
B1	-2.99	-5.96	1.22	0.13	2.66	0.01
H*	-2.7	-	1.49	-	2.65	-
S1	-2.24	-3.19	1.28	0.99	1.13	0.69
B2	-2.24	-2.08	1.05	1.07	1.85	1.87
P	-2.04	-2.3	1.15	1.02	1.79	1.46
A	-1.86	-2.3	1.27	1.23	1.03	1.15
B3	-1.69	-2.3	0.98	0.69	2.97	0.36
M7	-1.69	-2.3	1.15	0.67	1.99	0.39
C	-1.53	-2.55	1.33	1.14	1.88	3.06
S2	-1.38	-1.88	1.29	1.03	1.31	1.03
B4	-1.38	-1.7	1.1	0.95	2.45	6.78
L2	-1.38	-1.36	1.03	0.57	2.49	0.37
J1	-1.24	-0.91	1.36	0.91	1.74	1.26
L	-1.1	-1.06	1.08	0.64	2.53	0.47
Z	-1.1	-0.63	1.41	1.09	1.56	1.13
M1	-0.07	0.8	1.28	0.86	1.24	0.66
M2	0.05	1.07	1.48	0.91	1.55	0.71
M3	0.68	1.21	1.33	0.97	1.34	0.78
J2	1.22	2.41	1.46	1.06	1.88	0.92
N	1.22	2.9	1.5	1.01	1.9	0.51
E1	1.68	2.9	1.38	1.08	2.06	0.96
E2	1.68	2.9	1.57	1.12	1.87	0.76
M4	1.86	3.57	1.45	1.29	3.11	1.53
W	2.24	4.07	1.48	1.24	3.11	1.12
M8	2.72	4.85	1.12	0.8	3.42	0.12
J3**	3.01	-	1.43	-	4.28	-
M5	3.37	4.07	1.33	1.29	4.1	6.38
Mean	**-0.657**	**-0.434**	**1.279**	**0.957**	**2.314**	**1.640**
SD	**2.076**	**2.970**	**0.168**	**0.249**	**0.885**	**2.011**

*Pupil has obtained zero score, **Pupil has obtained perfect score

Table 3 presents paired t-tests analyses results. The sample is decreased to 28 pupils as two pupils after the changes in response patterns based on interview data obtained zero score and one pupil a perfect score. Statistical analyses (as reported by Table 3) showed that in relation to ability there was an average increase in ability estimates although this was not statistically significant (p=0.390). In relation to the fit statistics values, Table 3 shows that pupils' Infit values have decreased significantly ($p<0.0005$) after changing pupils' response

patterns based on interview data. Table 2 shows that overall the mean Infit value has decreased by approximately 0.30 units. This suggests that pupils' test behaviour as adjusted during interviews was more consistent with the measurement model. According to Tables 2 and 3, there was also an overall decrease noted for Outfit values but this decrease was not statistically significant (p=0.098).

Table 3: Paired Samples Test

		Paired Differences			t	df	Sig. (2-tailed)
		Mean	*Std. Deviation*	*Std. Error Mean*			
Pair 1:Ability	Before - After	-0.185	1.121	0.212	-0.873	27	0.390
Pair 2:Infit	Before - After	0.308	0.222	0.042	7.354	27	<0.0005
Pair 3:Outfit	Before - After	0.593	1.830	0.346	1.715	27	0.098

As significance testing analyses are affected by sample size we have also calculated the effect size for ability and fit analyses estimates so as to quantify the difference between the estimates obtained before and after changes in response patterns. Table 4 reports two statistics Cohen's d [1] and the effect size correlation[2]. Cohen's d (see Appendix for equation) reports the standardised difference between the mean estimates before and after the changes in response patterns while the effect size correlation is the point-biserial correlation. Cohen (1988) also suggested some general definitions for interpreting effect size estimates. Specifically according to Cohen an effect size is small if d =0 .20 or r =0 .10, medium if d = .50 or r = .30 and large if d = .80 or r = .50.

Table 4: Effect sizes for changes in ability estimates and fit statistics

	Cohen's *d*	**Effect size correlation**
Ability	0.073	0.037
Infit MNSQ	1.445	0.587
Outfit MNSQ	0.384	0.187

[1] **Cohen's $d = M_1 - M_2 / \sigma_{pooled}$** where $\sigma_{pooled} = \sqrt{[(\sigma_1^2 + \sigma_2^2) / 2]}$
Cohen (1988) defined d as the difference between the means (i.e. M1 - M2), divided by standard deviation (i.e. s), of either group. Cohen argued that the standard deviation of either group could be used when the variances of the two groups are homogeneous. In practice, the pooled standard deviation, pooled, is commonly used (Rosnow and Rosenthal, 1996).

[2] **$r_{Y\lambda} = d / \sqrt{(d^2 + 4)}$**

According to Cohen's benchmarks the effect sizes reported in Table 4, for ability and Outfit values are small while for Infit values very large. Cohen's benchmarks however reflect the typical effect sizes encountered in the behavioral sciences as a whole so when used in specific topic areas these may be misleading. Because of this limitation the mean estimates obtained before and after the changes in response patterns with 95% confidence intervals are also presented in Figure A1 (Appendix). Figure A1 shows that the difference in Infit estimates is striking. These findings suggest that a large part of unexpected performance as indicated by the Infit statistic can be accounted for by test modality. The same cannot be said though for unexpected performance as indicated by the Outfit statistic..

Although no overall significant changes were found for ability estimates, we wanted to examine the change in ability estimates for specific pupils as paired t-tests and effect size analyses referred to the change in the mean ability and not to specific pupils. It may be the case that certain pupils' ability estimates have been affected more than others. If we are interested to examine for cases of pupils who are seriously misrepresented by the test then we have to look for individual pupils whose abilities estimates have changed a great deal in either direction. In order to identify those pupils Table 5 was constructed.

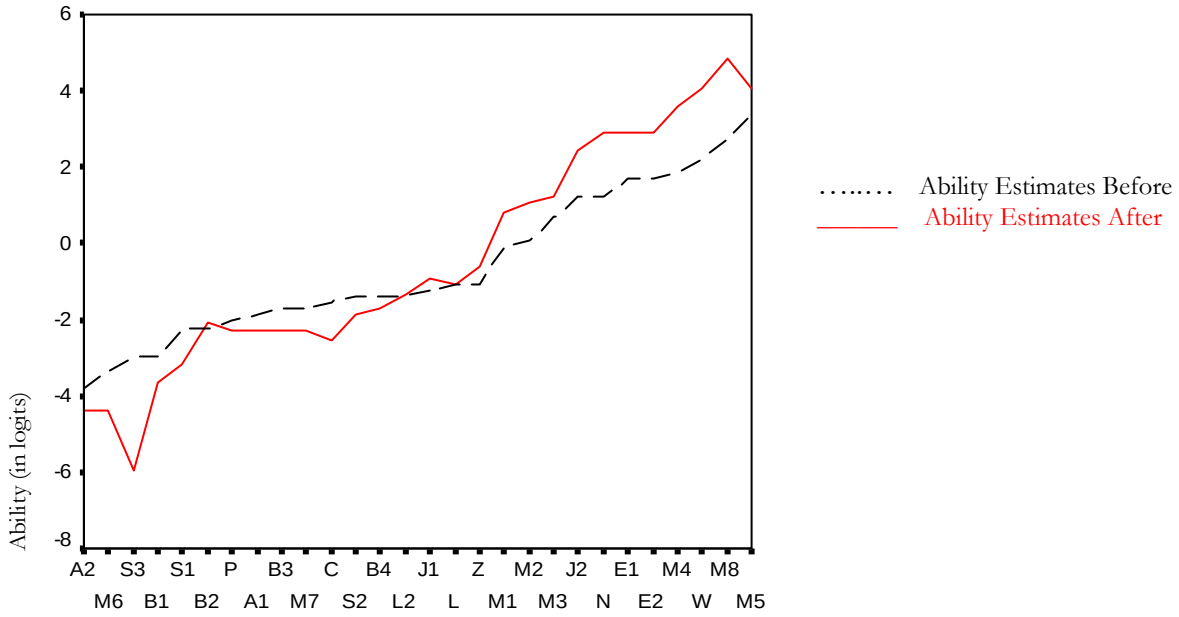
Table 5: Changes in ability estimates (in ascending order) for each pupil interviewed after the changes in response patterns based on what pupils showed able to do during the interviews

Pupil ID	Change in Ability Estimates (in logits)	Pupil ID	Change in Ability Estimates (in logits)
H	-∞	**B2**	0.16
T	-∞	**J1**	0.33
B1	-2.97	**Z**	0.47
M6	-1.03	**M3**	0.53
C	-1.02	**M5**	0.70
S1	-0.95	**M1**	0.87
S3	-0.66	**M2**	1.02
B3	-0.61	**J2**	1.19
M7	-0.61	**E1**	1.22
A2	-0.57	**E2**	1.22
S2	-0.5	**N**	1.68
A1	-0.44	**M4**	1.71
B4	-0.32	**W**	1.83
P	-0.26	**M8**	2.13
L2	0.02	**J3**	+∞
L1	0.04		

In Table 5 pupils interviewed have been ordered according to the change in their ability estimates. Positive changes signify that the ability estimates obtained after the changes in response patterns were higher than those originally obtained and vice-versa. Figure 1 presents graphically the results presented in Table 5 (i.e. ability estimates before and after the changes but presented in ascending order by 'ability before').

Both Table 5 and Figure 1 show that some pupils were seriously misrepresented by the test situation (e.g. B1, M8, W, M4 etc). It is obvious that these pupils have showed quite different pictures in the test and in the interview situation. This indicates that the test does not provide a representative picture of what these pupils are able to do orally in interviews. Moreover this also suggests that some pupils' ability estimates were greatly affected by the change in test modality. These cases require further action. The kind of action that should be taken depends heavily on the proposed use of these test scores and the decisions that these will inform.

Figure 1: Ability estimates before and after the changes in response patterns for each pupil interviewed (presented in ascending order by 'ability before')



The final set of analyses employed was simple regression in order to examine the effect of background variables on the change in ability estimates obtained. Specifically we wanted to examine what types of pupils were mostly affected by this change in test modality. For these

purposes the change in ability estimates became the response variable in a linear regression and four background variables were entered into the model i.e. initial Infit and Outfit values, gender (i.e. males=0; females=1) and language (Only English language spoken at home=0; an additional language other than English spoken at home=1). Table 6 presents the results.

According to Table 6 two variables had a statistically significant effect i.e. Infit and gender. These significant relationships suggested that males and pupils with higher Infit values had larger changes in their ability estimates. This indicates that males and pupils with high Infit values were the ones whose ability estimates were mostly affected by test modality. Specifically males and pupils with high Infit values obtained higher ability estimates after the changes in test response patterns. These significant relationships are presented also graphically in Figure A2 (Appendix).

Table 6: Parameter estimates reported from linear regression analyses

Response variable: Change in ability estimates	B	Sig.	R-squared	F-statistic
Constant	2.258	0.298		
Infit	2.389	**0.076**		
Outfit	0.072	0.771		
Gender	-0.889	**0.050**		
Language	0.293	0.563	0.442	4.522 (p=0.008)

Note. Numbers in bold indicate statistical significance.

Discussion

In this paper we examined the effect of assessment mode on ability estimates and specifically the stability of ability estimates as obtained in the test subject to change in 'test modality'; essentially the second 'mode' was 'test' and the second was 'test modifies by supplementary interview data'. In order to do this pupils' test response patterns were changed based on their responses during interviews. Fit analyses were re-run in order to obtain new estimates of their abilities and of the fit statistics. When old and new estimates were compared using a paired t-test, there was an overall increase in ability estimates and an overall reduction in pupils' Infit and Outfit values. However, only the change in the Infit values was statistically significant. Specifically, pupils' Infit values have decreased significantly after changing pupils' response patterns based on interview data. This finding indicates that a large part of

unexpected performance as indicated by the Infit statistic was accounted for by test modality. To our surprise the same was not true for unexpected performance as indicated by the Outfit statistic. Currently we are unable to explain this difference observed between the Infit and Outfit statistic. When interpreting the above findings one however has to keep in mind also that in the interview situation not all of the items in the test were administered but only items on which statistically unexpected responses were obtained. We are not aware whether pupils' responses on the rest of the items would have changed or not in the interview situation.

Although there was no significant change in the mean ability by examining the change in ability estimates for each individual pupil we have found cases of pupils for whom the two measures were seriously discrepant. It was obvious that some pupils have showed quite different pictures in the test and in the interview situation. This provides evidence that the test did not provide a representative picture of what these pupils are able to do in class and also raises serious concerns about the validity of test scores for these individual pupils. These cases require further action and the kind of action to be taken depends heavily on the purposes and proposed uses of the test. In case of high-stake tests this would result to serious adverse consequences for the pupils and institutions involved (potentially casting doubt consequential validity).

The above findings suggest also that some pupils' ability estimates were greatly affected by the change in test modality. By regression analyses we have found that the pupils whose ability estimates were mostly affected by test modality were males and pupils with highest Infit values, but interestingly not second language.

Conclusion

In this study we have shown that some pupils' (both low and high ability) ability estimates were greatly affected by the change in test modality. This finding raises serious concerns about the validity of test scores for these individual pupils. Another significant result is that overall Infit values decreased significantly after changing pupils' response patterns based on interview data. For Outfit values there was also an overall decrease but this decrease was not found to be statistically significant. These findings indicate that a part of unexpected performances can be accounted for by test modality. This pattern appears to be more

evident in the case of the Infit statistic where a large part of the unexpected performances was accounted for by test modality. While this result is promising, further work needs to be done to explore this effect, in particular to defend this from threats to validity such as 'regression to the mean', caused by the method used here, i.e. only interviewing examinees about unexpected responses.

References

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person-fit. *Applied Psychological Measurement*, 25(2), p. 107-135.

Petridou, A. & Williams, J. (under review). Accounting for Real Person Misfit using Multilevel Models. (Revise and re-submit for the *Journal of Educational Measurement*)

Petridou, A & Williams, J. (2006). Explaining Examinee Misfit: pupils' and teachers' explanations. Paper presented at the Annual Conference of American Educational Research Association. San Francisco.

Rosenthal, R. & Rosnow, R. L. (1991). Essentials of behavioral research: Methods and data analysis (2nd ed.). New York: McGraw Hill.

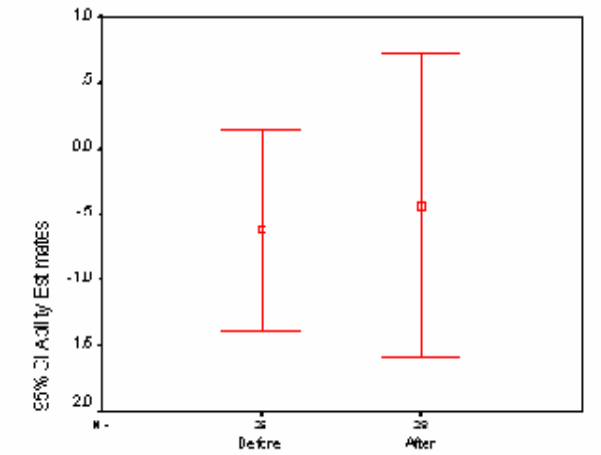
Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, 46, p. 359-372.

Smith, R. M. (1987). Theory and practice of fit. *Rasch Measurement Transactions*, 3:4, p. 78.

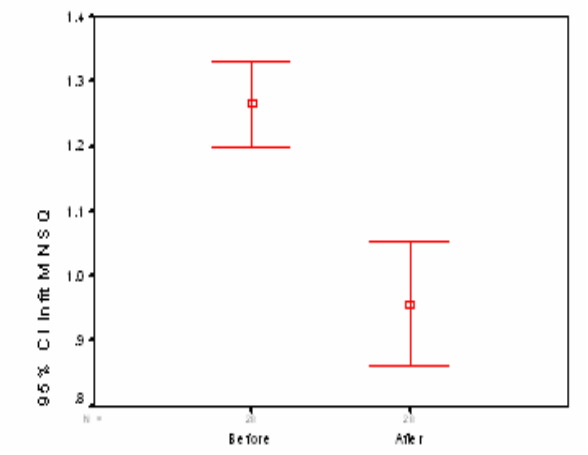
Appendix

Figure A1: Mean Estimates with 95% Confidence Intervals before and after the changes in response patterns

Ability Estimates



Infit MNSQ



Outfit MNSQ

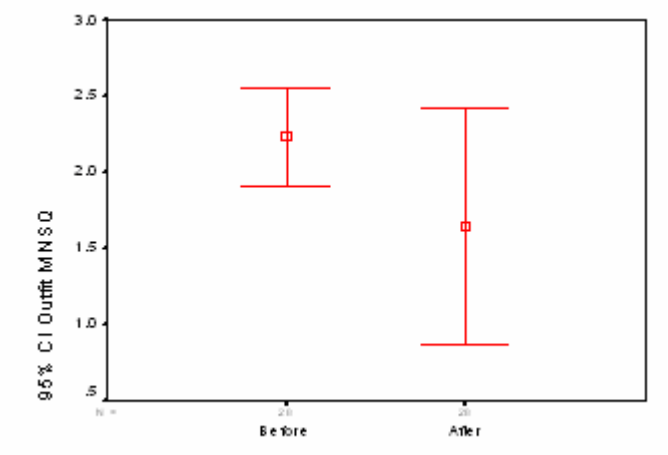


Figure A2: Graphical presentation of significant relationships in simple regression