Psychology Science Quarterly, Volume 50, 2008 (3), pp. 433-447

Measuring change in training programs: An empirical illustration

RENATO MICELI[1], MICHELE SETTANNI[1] & GIULIO VIDOTTO[2]

Abstract

The implementation of training programs often requires a complex design if effectiveness is to be accurately evaluated. Part of the difficulty lies in the fact that trainees must be presented with a series of ever-changing tasks in order to avoid biases due to learning or carryover effects.

The aim of the present study is to experiment and illustrate a simple procedure, based on a special case of the linear logistic test model (LLTM), used to evaluate the effectiveness of a training program. The procedure is empirically applied to a dataset derived from a moped riding skills training program. The sample is composed of 207 high school students who took part in three training sessions using a riding simulator. A different task presentation order was assigned to each subject and the whole design was completely balanced. The procedure applied allowed us to obtain estimates of the overall change in ability that occurred over the course of the training process. Furthermore, we were able to obtain estimates of item and subject parameters unbiased by the influence of change in ability due to training. Implications of the results are discussed and suggestions for future research are presented.

Key words: Measuring change, Linear Logistic Test Model (LLTM), Multi-Facet Rasch model, Training effectiveness

[1] Renato Miceli, email: renato.miceli@unito.it; Michele Settanni, email: settanni@psych.unito.it. Both: Department of Psychology, University of Torino, Via Verdi 10, 10124 Torino, Italy, Europe.

[2] Giulio Vidotto, Department of General Psychology, University of Padua, Via Venezia 8, 35131 Padova, Italy, Europe; email: giulio.vidotto@unipd.it

Introduction

Measurement of change and Item Response Theory

The measurement of change is an important issue in many social science contexts. Over the past several decades psychometricians and statisticians have attempted a number of different approaches to quantifying change due to development, learning, or other events. Bereiter (1963) highlighted some fundamental problems derived from a naïve approach to the issue.

One of the main problems in the study of change using test scores is the lack of an interval level of measurement, as equal numerical changes in test scores do not represent equal affective or cognitive changes in latent ability at different levels of the continuum. The inequality of such changes is particularly evident when floor and ceiling effects manifest themselves in pre- and post-testing. The disadvantages of these effects are described by Fischer (1976).

One possible method for overcoming problems tied to unequal measurement intervals is to use those models of latent trait theory which characterise the qualitative responses of persons to test items in terms of person and item parameters (Kissane, 1982). A simple and convincing model for these purposes is the Rasch model, which considers a single ability parameter for each person, and a single difficulty parameter for each item. The advantage of the Rasch model over other measurement models is that no distributional assumptions about either person or item parameters need to be made (Rasch, 1960). In addition, by being able to choose different subsets of items which conform to the model, different tests of varying difficulties can be linked in order to provide much broader tests, thus overcoming floor and ceiling effects (Wright, 1977).

One area in which measurement of change is essential is the evaluation of training effectiveness. Training contexts are excellent examples of situations in which the measurement of change is required, in particular to evaluate the effectiveness of the training procedure. Some training procedures may actually result in no improvement in the ability for which the training was implemented, or worse yet, may even cause a significant decrease in the ability level it was intended to improve.

One of the most promising approaches to the issue of change measurement is Item Response Theory and, more specifically, models derived directly from the original Rasch model (Rasch, 1960). One of the first authors to address this issue was Fischer (1973), who developed the linear logistic test model (LLTM), an extension of the original Rasch model applicable in measurement contexts in which dichotomous items are administered more than once to the same individuals. In order to measure change, LLTM requires that two conditions are met: 1. the test employed must be composed of a set of k unidimensional items (displaceable on the same latent continuum); and 2. all items must be presented at all time points (Glück & Spiel, 1997). It tests the significance of effects and of differences between effects in experimental or quasi-experimental designs and explains differences in difficulty between items by modelling the difficulty parameters as linear combinations of some basic parameters. With regard to the measurement of change, the difficulty parameter of a test item at *t2* is assumed to be the result of the sum of the item difficulty at *t1* plus one or more parameters accounting for the change (Glück & Spiel, 1997). In this model, it is assumed that the item difficulty parameter of the Rasch model, σ_i, can be decomposed into the difficul-

ties of a set of basic parameters smaller than the number of items, e.g. it allows a set of structural parameters to be broken down into a weighted sum of component parameters:

$$p(X=1\,|\,\theta_v,\sigma_i)=\frac{\exp\left[\theta_v-(\sum_{j=1}^{J}q_{ij}\eta_j+c)\right]}{1+\exp\left[\theta_v-(\sum_{j=1}^{J}q_{ij}\eta_j+c)\right]} \tag{1}$$

where X indicates the response of person *v* on item *i*, θ_v is the ability parameter of person *v*, η_j is the difficulty parameter of component *j*, the q_{ij} denote the a priori fixed constants that define the weight of component *j* for item *I,* and *c* is a normalisation constant.

Fischer later expanded his approach elaborating the Linear Logistic Model with Relaxed Assumptions (LLRA). The LLRA (Fischer, 1976, 1977, 1983, 1987, 1989, 1996; Fischer & Formann, 1982) was specifically developed for measuring change. Through the use of virtual subjects LLRA allows group-specific changes to be estimated even when test items do not measure the same latent dimension (Glück & Spiel, 1997).

More recently, Linacre (1994) elaborated a model based on LLTM (for a more detailed explanation of its connection to LLTM see: Glück & Spiel, 1997; Rost & Carstensen, 2002) but specifically addressed to deal with tests designed according to a facet structure. This model is called Multi-Facet Rasch model, abbreviated as MFRM. Linacre's aim was to develop an effective way to include in the measurement process the effect of "facets" in addition to the two main facets (item difficulty and subject ability) considered by the Rasch model, without losing its measurement properties.

In a facet-designed test, each item can be seen as the result of a systematic combination of two or more factors or facets. These facets may represent different contexts or time points when the items were administered, or they can account for the effect of different judges evaluating responses to the items. These additional facets may be regarded as the decomposition of the original single Rasch item difficulty parameter (Linacre, 1994). Linacre assigned specific sets of parameters to each facet so that the response probability could be modelled as the logistic function of the sum of an ability parameter, θ_v, plus a content parameter, σ_i, plus a situation parameter, δ_t. Hence, the probability a person *v* giving a correct answer to an item *i* in the situation *t* can be described as:

$$p(X=1\,|\,\theta_v,\sigma_i,\delta_t)=\frac{\exp(\theta_v-\sigma_i-\delta_t)}{1+\exp(\theta_v-\sigma_i-\delta_t)}. \tag{2}$$

One of the clearest examples of how a different facet can affect the measurement process is represented by the severity of raters (or judges) in sport competitions. For instance, in the case of sports such as gymnastics or diving, it is evident that the final result of the athlete's performance (the final score) depends not only on the difficulty of the task and on his/her ability, but also on the severity of the judges. Linacre developed the MFRM in order to allow for examination of such, or even more complex assessment situations, e.g. to assess the influence of more than three factors. The multi-facet model allows all the relevant facets of a measurement situation to be modelled concurrently but examined independently (Bond &

Fox, 2001). Each facet is calibrated conjointly from the observed ratings. A person's ability is estimated based on all ratings given by all judges on all items. Judge severity is estimated based on all ratings given across all persons and items; and so on. This makes it possible to estimate the locations of persons, judges, and items on a common interval scale, which represents the frame of reference for understanding the relationships with all the facets of the measurement context.

Another important consideration is that, while each subject must usually face all items, very often each judge does not rate the performance of every subject. For this reason, in order to make the model estimable, assuring a sufficient connection between subjects and items, the measurement situation must be designed accordingly (Linacre, 1994).

When dealing with tests including polytomous items, the Multi-facet version of the Partial Credit Model (Wright & Masters, 1982) can be written as follows:

$$p(X=1 \mid \theta_v, \sigma_i, \delta_t, \tau_{ik}) = \frac{\exp(\theta_v - \sigma_i - \delta_t - \tau_{ik})}{1+\exp(\theta_v - \sigma_i - \delta_t - \tau_{ik})} \tag{3}$$

where τ_{ik} is interpretable as the additional difficulty needed by subject *v* to reach level *k* of item *i*.

The MFRM estimation equations derived by Linacre (1994) are in a way similar to those obtained for the polytomous models by Wright and Masters (1982), using unconditional maximum likelihood (Fisher, 1922). These equations yield sufficient parameter estimates and asymptotic standard errors for the ability of each subject, the difficulty of each item, the severity of each judge, and the additional level of performance represented by each step on the partial credit scale.

In the present work we employed the MFRM to model the influence of a different kind of factor: time. The basic idea is that subsequent administration of test items to the same subjects (with items presented in different order according to a precise experimental design), as occurs during training based on prolonged experience with similar tasks, can be modelled using Linacre's model. In such a situation we can conceptualize a subject's performance as being affected by a third factor in addition to the subject's ability and item difficulty: cumulative experience. It is logical to expect that a trainee's ability will improve the more he or she practices a training task. Essentially, it is the subject's ability that is likely to improve, but this process can be better described by decomposing ability in two different components: true subject ability and the effect of prolonged experience (or time). In this way, it is possible to assess how the trainees' ability evolves over the course of the training.

Fit statistics, reliability, and separation index

The multi-facet Rasch model also provides estimates of the consistency of the observed response patterns. The fit analysis is based on the computation of the difference between expected and observed scores.

Unexpected low ratings for able subjects are improbable failures, often due to careless mistakes or misunderstanding, while unexpected high ratings obtained by scarcely able subjects are improbable and usually due to lucky guessing or special knowledge.

Two fit statistics have been used here: *Outfit and Infit* (Wright & Masters, 1982). They can be computed for each of the facets involved in the measurement process, however their computations will be shown for items only.

Rasch models supply a direct estimate of the modelled error variance for each estimate of a person's ability and item difficulty (Wright, 1999; Wright & Masters, 1982; Wright & Stone, 1979). The same holds true for MFRM as well (Linacre, 1994). Individual standard errors (*SE*s) are more useful than a sample or test average, which overestimates the error score variance of persons with high and low scores. The Rasch measurement model is able to produce an optimal estimate of internal consistency because the numerical values express interval scale measures if the data fit the model, and the actual average error variance of the sample is used instead of the error variance of an average person. Based on these considerations, Wright and Stone (1979) developed person separation reliability (R_{sep}), which is an index of the sample standard deviation in terms of standard error.

The separation index (also referred to as *G* index) can be defined as the ratio of the standard deviation of the sample expressed in logits adjusted for inflation due to error (i.e., true variance) to the standard error of measurement. The ratio expresses the relationship between the amount of variability (standard deviation adjusted for error) within the sample to the precision (Root Mean Square Error, RMSE) with which that variability was measured. It is a measure of how well the instrument's items separate the subjects in the sample (Wright & Masters, 1982).

The higher the value of *G* (and hence R_{sep}), the more spread out the subjects are on the variable being measured (Schumacker & Smith, 2007).

Purpose of the study

The present study has been carried out with the purpose of testing a simple procedure based on the Multi-facet Rasch model to evaluate the effectiveness of a training program. Procedure and outcomes will be presented and discussed employing empirical data derived from training sessions with a riding simulator.

Materials and methods

Sample

The sample consists of 207 high school students aged 14-15, balanced for gender (51% females), living in northeast Italy. The main subject inclusion criterion was, in addition to age, that the students would have begun riding a moped in the next six months.

Motivation to take part in the project was supported by a reward, which was the opportunity to earn school credits that could be used by the subjects in their final examination. Formal consent to participate in the research was obtained from all the students' parents.

Riding simulator and training procedure

The data derives from a training programme for improving riding ability through the use of a riding simulator (Honda Riding Trainer, HRT). The simulator was developed as a means to improve drivers' hazard awareness, coordination, and perception skills by allowing them to safely experience hazards in a variety of settings (e.g. in a city, on a motorway). A number of training situations are included with various training modes in differing environments.

The HRT has been set to propose twelve full tracks, including hazard scenes, automatic replay, and a final summary: six tracks are located in a city with wide streets (tracks P_01 to P_06); five tracks are located in a city with narrow streets (tracks S_01 to S_05); one track is located in a residential area (track T_01). Each track consists of eight hazard scenes. All hazard scenes come from a European study that analysed in detail more than one thousand road accidents involving motorbikes and mopeds (MAIDS, 2004).

Data collection

The HRT collects data for each subject during the training sessions. Data are obtained from two sources: the handlebar device (e.g. accelerator activation, brake activation, handlebar turn angle, use of turn signals, peripheral view, and gear position) and HRT internal variables.

For the aims of the present study, only internal HRT data were used. In particular, analyses were carried out using scores provided by the HRT on the performance of every subject facing each hazard scene presented during the training. The internal software is programmed to assign a grade for each subject's performance. The four possible grades are "D" for accident, "C", "B", and "A" respectively for sufficient, good, and excellent performance. The different "non-accident" grades are intended to reflect the ability of the trainee in avoiding the risk proposed by the scene: the closer the accident (due to high speed, abrupt stops, or lane changes, etc.) the lower the assigned grade.

A different track presentation order was assigned to each subject and the whole design was completely balanced, as shown for a subsample in Table 1, in order to have all the tracks, and consequently all the hazard scenes, presented in every possible order (for a more detailed presentation of data collection and further different analyses see Settanni, 2008).

Table 1:
Presentation order of the tasks (for a subsample)

		Track presentation order											
		[, 1]	[, 2]	[, 3]	[, 4]	[, 5]	[, 6]	[, 7]	[, 8]	[, 9]	[,10]	[,11]	[,12]
Subjects	[1,]	2	4	11	8	14	1	10	6	13	3	7	5
	[2,]	8	10	3	14	6	7	2	12	5	9	13	11
	[3,]	5	7	14	11	3	4	13	9	2	6	10	8
	[4,]	1	3	10	7	13	14	9	5	12	2	6	4
	[5,]	3	5	12	9	1	2	11	7	14	4	8	6
	[6,]	12	14	7	4	10	11	6	2	9	13	3	1

Data analysis

The grades assigned by the simulator to the subjects (one per hazard scene) were arranged in a rectangular matrix (subject x item) to allow for analysis using statistical software.

The first operation performed was the screening of the frequency distributions of the grades for each item. This was required in order to avoid the presence of items with void response categories, i.e., to exclude from analysis any item for which the HRT was unable to assign a grade.

The following step was an analysis of the score data obtained from HRT based on the Multi-facet Rasch model. This enabled us not only to estimate parameters referring to item difficulty and subjects' ability, but also to estimate parameters related to the influence of track presentation order. This analysis was carried out to assess the indirect influence of the training on subject performance. To make this step clearer, it is useful to draw a comparison between the effects of the presentation position and the influence of judge severity in a more typical test situation. Each possible track presentation order (from the first to the twelfth) can be conceptualized as a different judge evaluating subject performance. The tracks presented earlier in the training are expected to be evaluated by stricter judges. The greater the number of tracks encountered, the more lenient the judge. Clearly our expectation was to find a general decreasing trend in "judge severity", actually indicating a symmetrical increase in subjects' ability over the course of the training.

Facets software (Linacre, 2004) was employed for the MFRM analysis, which allowed measures to be constructed from complex data involving heterogeneous combinations of examinees, items, tasks, and judges as well as other measurement and structural facets. *Facets* utilises the estimation algorithm JML (de Jong & Linacre, 1993).

Results and discussion

Preliminary screening

The first step of the analysis was the screening of the frequency distributions of the grades (response categories) assigned by the HRT simulator to the participants.

The aim of this step was to assess the possibility of retaining all the possible grades in the subsequent analyses.

Many of the items were found to have empty response categories, meaning that for those items any participants got one or more of the possible grades. This issue had to be confronted in order to prevent potential problems in parameter estimation. For this reason, the number of possible grades was reduced from four to three, collapsing the central grades into one single rating, corresponding to sufficient or good performance. This allowed for improved response (grade) distribution for many of the items, though some problems did persist: Even after the recoding, some of the items still had empty categories. Given the possible bias associated with the presence of such items, they were excluded from further analyses. A total of nine items were excluded.

Multi-facet Rasch model analysis

In the following step of the analysis, the MFRM was applied to estimate item and subject parameters, meanwhile considering the effect of the order of presentation of the tracks, i.e. the effect of cumulative experience with the simulator on subject performance. More precisely, it was expected that, on the first tracks, subjects would have performed more poorly than on subsequent ones, due to improved ability. This expected gradual change in ability may be computed and described as the parameter δ_t (equation 3) of the MFRM. Consequently, the estimates regarding the third facet of this model (e.g. presentation order) represent the effect on performance of progressive experience with the simulator and may be interpreted as change in ability caused by practice with the simulator.

If HRT actually improved the trainees' level of driving ability, then the estimated measures of the twelve subsequent presentation orders should represent the change in driving ability that occurred during (and because of) the training.

Misfit diagnosis

Item and person fit statistics. The first step in the MFRM analysis was to consider the fit of both items and participants. The employed software allowed us to compute fit statistics for subjects, items, and presentation order. The rationale behind the interpretation of these statistics is to be able to assess the underlying assumptions about dimensionality and local independence that must accrue for invariant, equal-interval scaling.

Study of item fit provides insight into whether all the items sample the same underlying trait or at least a set of underlying personal factors that function together to determine the subjects' performance in the same way on each item (Bond & Fox, 2001) and is central to the validity of an assessment tool. If an item does not produce ratings that fit the pattern expected according to the MFRM it is assumed that the item is not measuring the same construct as other items. For instance, it could be poorly written (in this case grade assignment might be poorly "programmed"), or raters might interpret a rating scale differently from the way it was intended by test developers, consequently producing unexpected responses.

With regard to participants, fit statistics have been computed in order to detect individuals whose response patterns do not fit the pattern predicted by the MFRM. The presence of such individuals, indeed, decreases the precision of difficulty parameter estimates.

According to Linacre (2002) an acceptable range for Infit and Outfit values is between 0.5 and 2.0. Infit or Outfit values of less than 0.5 indicate that the element of the facet (item, person, or other) does not provide information beyond that provided by the rest of elements on the scale. For instance, this can occur when there are several items that are similar or highly correlated or when one item is dependent on another. In contrast, Infit /Outfit values greater than 2.0 indicate that the element does not define the same construct as defined by the rest of the elements. With respect to items, it could mean that the item is either poorly constructed or misunderstood, or that it is ambiguously defined. Items with such values may distort or degrade the measurement system. However, even items with Infit/Outfit values between 1.50 and 2.00 should be examined carefully as, though not necessarily degrading, they may be unproductive for construction of measurement.

With regard to the HRT items, our analysis yielded the following results: The maximum Infit value detected was 1.40 (item S_05.08) and there were no overfitting items (Infit <.50), with minimum Infit value for item S_01.07 (Infit = .78).

The Outfit values showed the presence of only one slightly misfitting item, T_05.08, with Outfit = 2.35, but with a corresponding Infit value in the acceptable range (Infit = 1.03). In terms of Outfit values as well, no overfitting items were detected.

With respect to trainee fit, looking at Infit, only one misfitting individual was detected with an Infit value of 1.85. No overfitting persons were found.

In terms of Outfit values, eight participants were found with Outfit values greater than 1.50 but none of them had a value exceeding 2.00. Furthermore, the highest Infit value was linked to the participant with the highest Outfit value. Thus, this person's response pattern does not appear to fit the model and the usefulness of his/her presence in the data set is questionable. However, for this study, the subject was retained for the analyses.

Presentation order fit statistics. We used *Facets* to compute fit statistics for the twelve presentation orders as well. This allowed us to evaluate the consistency of the effect of the different presentation order on subject performance. Fit statistics are reported in table 2.

Table 2:
Presentation order fit statistics

Presentation order	Infit	Outfit
#1	1.05	1.05
#2	1.07	1.03
#3	1.04	1.01
#4	1.00	1.09
#5	0.94	0.97
#6	0.88	0.84
#7	1.03	0.97
#8	0.98	0.93
#9	1.08	1.14
#10	1.01	0.91
#11	0.99	1.05
#12	0.99	0.98

As seen in the table above, orders of presentation fit the MFRM well. There were no presentation orders with Infit or Outfit values outside the acceptable range. Indeed, Infit ranges from .88 to 1.07, while Outfit values are all in the range of .84 – 1.14. The absence of fit problems concerning the presentation order facet indicates that training experience, in this case operationalised as the number of sessions already completed, had a consistent effect on trainees' ability.

Reliability and separation index

Reliability values for items, subjects, and presentation orders are shown in table 3. High person reliability means that we have developed a test on which some persons score higher and some lower and that we can trust the consistency of these inferences. The value of person reliability computed by *Facets* is .89, indicating good replicability.

Even with respect to items, the value of reliability found was very high (r_{sep} = .98). Hence, we can have confidence in the item difficulty estimates.

Unlike the person reliability estimate, which has a maximum value of 1.00, the person separation index, *G*, is not constrained by an upper boundary, but has a range of zero to infinity. The recommendation is that the separation ratio should exceed 2; in other words, the variability in the sample should be at least twice the variability of noise in the test. It was 2.81 for this sample, suggesting more than acceptable reliability.

Table 3:
Reliability and separation indexes

Facet	Reliability (R_{sep})	Separation index (G)
Subject	.96	2.81
Item	.98	7.76
Presentation order	.98	7.80

Subjects and item estimates

Figure 1 depicts measures of subjects and items on the same map. Inspection of the subject-item map indicates that the items are not well targeted toward the higher end of driving ability. This result seems to indicate that, on the whole, the training sessions were experienced by subjects as quite easy, i.e., for most of the subjects the majority of the hazard scenes were easy to deal with. Table 4 shows summary item and subject statistics.

Item measures range between -2.19 (item P_01.01) and 3.54 (item S_03.08) with an average standard error of .14.

With regard to the subjects, their ability measures are comprised between -.81 and 3.02 logits, and are less dispersed than the item measures (*SD* = .57). Mean standard error is .14 and the average measure is .97, almost 1 logit higher than the mean of item difficulty. The items are generally easy for the participants. As seen in table 3, the accuracy of the measures (which is different at different levels of the continuum) is generally high both for items and subjects. Expressing these values as mean reliability, we found a value equal to .96 (min = .90) for subjects and .98 (min = .85) for items.

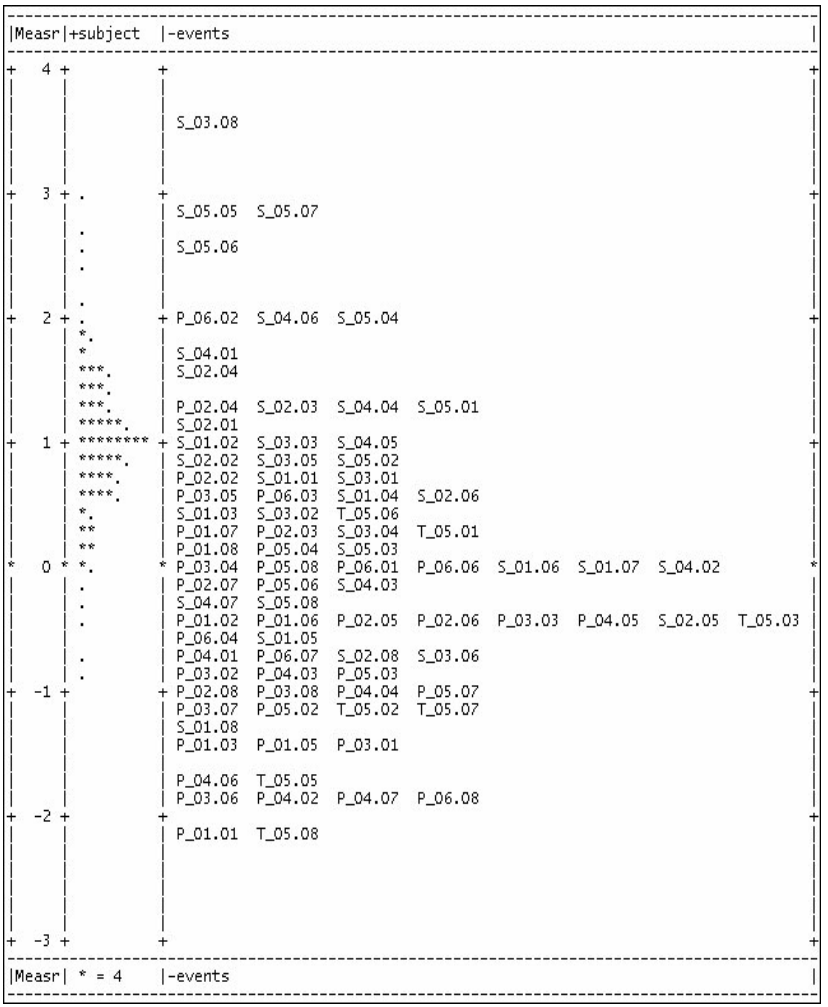

Figure 1:
Map of items and subjects based on MFRM calibration.

Table 4:
Summary statistics

	Participants (ability)	**Items (difficulty)**
N	207	86
Mean	.97	.00
SD	.57	1.18
Min	-.81	-2.19
Max	3.02	3.54
MSE[SD; min; max]	.20 [.02; .18; .31]	.14 [.05; .09; .39]

Presentation order estimates

The main reason for having employed MFRM in this research is that it allowed us to estimate measures related to different presentation order. These estimates, as explained previously, can be interpreted as change in subject ability due to practice with the simulator.

Figure 2 shows the estimates for the subsequent training sessions (standard errors ranging between .04 and .05). As the figure clearly shows, employing the MFRM made it possible to detect a significant influence on performance attributable to the training itself. Performances tend to improve with the training and this effect can be accounted for by practice with the simulator. This result can be interpreted as an increase in the ability of the participants corresponding to the length of time spent using the instrument.

More specifically, the steepest line is the one connecting the first and the second track. This might indicate that after the first items, i.e. after having encountered the first hazard scenes, trainees had learned how to use the simulator controls correctly. Subsequently, there was constant improvement in performance, with the exception of small decreases corresponding to the first track of both the following training sessions, probably due to the process of becoming re-acquainted with the simulator. At the end of the curve, a plateau can be observed, which probably indicates achievement of the maximum level of improvement in driving ability attributable to the HRT.

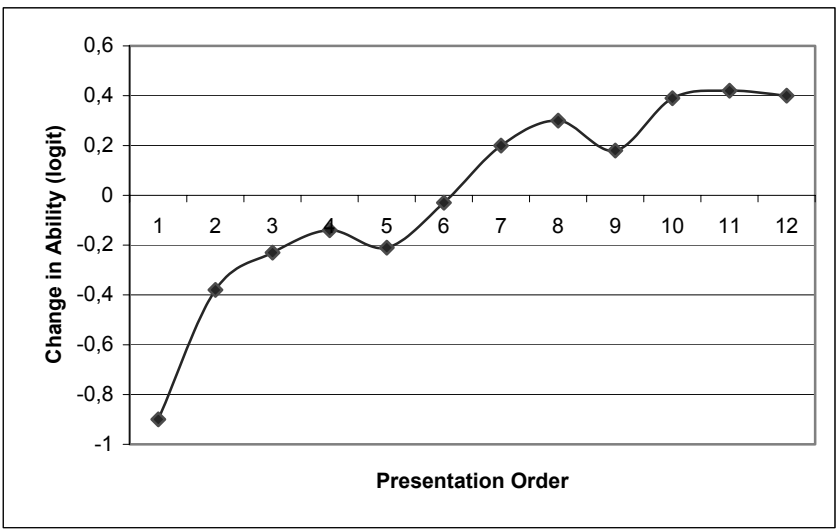

Figure 2:
Presentation order estimates.

Conclusions

The procedure described here allowed for the application of a special case of LLTM to a training programme, with a structure that is quite common in real life, i.e. the subsequent administration of ever-changing items. Use of the Multi-facet Rasch model made it possible to obtain estimates of overall change in ability due to the training over the course of the training process. This allowed us to monitor not only the potential presence of a substantive change caused by the entire training procedure, e.g., confronting initial and final estimates, but also to detect, if present, particular or local trends of induced change, making it possible to identify unexpected or flawed functioning of the training programme.

At the same time, having partialled out the learning process, estimates were obtained of item and subject parameters free from the effect of the change in ability due to training. In particular, estimation of the presentation order facet freed the difficulty parameters from the influence of their position in the test, which allowed item difficulty parameters to be obtained which were not biased by order effect.

On the whole, the use of the MFRM as proposed here was found to be useful in assessing overall training effectiveness. However, a major issue which was not addressed here is the stability of interindividual differences over time or, more precisely, the consistency of interindividual differences in intraindividual change. Given the fundamental importance of this aspect for the study of training effectiveness, further research is needed (and is actually in progress by the authors) to find an effective way to obtain individual measures of change.

Acknowledgements

The data employed in this study was collected as part of a research project carried out by the Department of General Psychology of the University of Padua in collaboration with Honda Europe, who partially supported this work. We would like to thank Honda Motor Europe for having allowed us to use the data.

References

Bereiter, C. (1963). Some persisting dilemmas in the measurements of change. In C. W. Harris (Ed.), *Problems in the measurement of change* (pp. 3-20). Madison, WI: University of Wisconsin Press.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: fundamental measurement in the human sciences,* London: Lawrence Erlbaum Associates.

de Jong, J., & Linacre, J. M. (1993). Estimation methods, statistical independence and global fit. *Rasch Measurement Transactions, 7*(2), 296-297.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359-374.

Fischer, G. H. (1976). Some Probabilistic Models for Measuring Change. In D. N. M. de Gruijter & L. J. Th. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement* (pp. 97-110). London: John Wiley.

Fischer, G. H. (1977). Linear logistic models for the description of attitudinal and behavioral changes under the influence of mass communication. In W. F. Kempf & B. H. Repp (Eds.), *Mathematical models for social psychology*. Bern: Huber, 1974/ New York: Wiley, 1977.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3-26.

Fischer, G. H. (1987). Applying the principles of specific objectivity and generalizability to the measurement of change. *Psychometrika, 52,* 565-587 .

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika, 54*, 599-624.

Fischer, G. H. (1996). The linear logistic test model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 225–243). New York: Springer-Verlag.

Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement, 4,* 397-416 .

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical. Transactions of the Royal Society of London (A) 222*, 309-368. [Reprinted as Paper 18 in Fisher, 1974].

Fisher, R. A. (1974). *The collected papers of R. A. Fisher* (ed. J. H. Bennett), Adelaide: University of Adelaide Press.

Glück, J., & Spiel, C. (1997). Item response models for repeated measures designs: Application and limitations of four different approaches. *MPR-online, 2, 1*. Retrieved from: http://www.dgps.de/fachgruppen/methoden/mpr-online/issue2/art6/glueck.ps

Kissane, B. V. (1982). The Measurement of Change as the Study of the Rate of Change. *Education Research and Perspectives, 9*, 55-72.

Linacre, J. M. (1994). *Many-Facet Rasch Measuremen* (2nd Ed). Chicago: MESA Press.

Linacre, J. M. (2002). What do *Infit* and *Outfit*, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2004). *Facets Rasch measurement computer program.* Chicago, IL: Winsteps.com.

MAIDS (2004). *In-depth investigation of motorcycle accidents: The first complete European in-depth study of motorcycle accidents. Final Report 1.2.* Retrieved June 1, 2006, from http://maids.acembike.org.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute of Educational Research. (Expanded ed., 1980. Chicago: The University of Chicago Press)

Rost, J., & Carstensen, C. H. (2002). Multidimensional Rasch Measurement via Item Component Models and Faceted Designs. *Applied Psychological Measurement, 26,* 42.

Schumacker, R. E., & Smith, E. V. (2007). Reliability: A Rasch perspective. *Educational and psychological measurement, 67*(3), 394-409.

Settanni, M. (2008). Development of a Rasch-based method for the evaluation of training effectiveness: An empirical illustration. Unpublished doctoral dissertation. University of Torino.

Wright, B. D. (1977). Solving measurement problems with the Rasch model [Electronic version]. *Journal of Educational Measurement, 14*(2), 97-116.

Wright, B.D. (1999). Fundamental Measurement for Psychology. In S.E. Embretson & S.L. Hershberger (Eds.), *The New Rules of Measurement: What every psychologist and educator should know* (pp. 65-104). Mahwah, NJ: Lawrence Erlbaum Associates.
Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*, Chicago: MESA Press.
Wright, B. D., & Stone, M. H. (1979). *Best Test Design*, Chicago: MESA Press.