**Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione**

# WORKING PAPER N. 29/2016

## Examining the relationship between items position and their functionality

Clelia Cascella – INVALSI, clelia.cascella@invalsi.it

**Abstract**

The investigation of the impact that items administration order within an achievement test (may) have on their functionality has received an increasing attention in last decades. This is a relevant topic in test development, first at all, because item position can bias estimated item parameter and it would ultimately result in biased ability estimation too, causing an unfair scoring for examinees. It is relevant also because possible differences in items parameter estimation could be explained taking into account that some students might have an advantage working on a certain item at a certain position, while the other might be handicapped. For example, if there are different sequences of item presentation, different solving strategies (i.e. based on different cognitive processes) might be activated by students.

In order to control if and how items administration order affect both items and, consequently, overall achievement test functionality, the Rasch Model was applied to analyze answers given by 532 pupils attending the $5^{th}$ grade level of primary school to 4 achievement tests administrated to measure mathematical ability.

Empirical results showed little differences in items parameter estimation depending on their location order. Nevertheless, some relevant differences were disclosed in scales construction process as well as in item functionality by means of Item Characteristic Curve. Differently from recent literature, these empirical evidences confirmed the existence of some effects of item position on their psychometrical properties.

**Keywords**

Item-position; Achievement test; Cognitive processes; Mathematics; Rasch model.

## Introduction

The impact of items administration order on their psychometrical properties has receiving an increasing attention, first at all because altering items position within the test is a normal practice in order to prevent cheating, in both paper-and-pencil and computer-based administration. In addition, currently, test designs very often contain several test booklets with the same items presented at different test positions. In fact, «(...) different examinees presented with different sequences of items are most likely not compared in a fair manner. This is due to the fact that one examinee might have an advantage working on a certain item at a certain position, while the other might be handicapped (...) Different sequences of item presentation occur systematically within large-scale assessments where various test booklets with partly different item subsets are used» (Hohensinn, *et al.*, 2008, 392).

Nevertheless, previous studies proved that no significant item-position effects can be disclosed when some conditions (such as, to have enough testing time) are guaranteed (Mollenkopf, 1950; Sax & Cromack, 1966; Flaugher, Melton, Myers, 1968; Hahne, 2008). On the other hand, other studies disclosed just a small fatigue effect and no other global and constant position effect (e.g., Hohensinn, *et al.*, 2008; 2011). The most important conclusion of those cited studies was that, «(...) all in all, as a consequence for future test designs, another order of these calibrated items should lead to basically unbiased (at least within the limits of standard error) person parameters. Of course, this is only true as long as the testing conditions are the same, in particular, as long as the number of items for each booklet and the testing time are the same. Nevertheless, comparing our results to findings of other studies reveals that position effects should be examined for every newly constructed assessment which deals with booklet designs» (Hohensinn, et al., 2011, 508).

Other studies reported different results. For example, in large-scale assessments with booklet designs, some item position effects (in particular, fatigue effect) were found for PISA (the Programme for International Student Assessment), in 2000; for data collected by OECD (the Organisation for Economic Co-operation and Development, in 2002); and, for TIMSS (the Trends in International Mathematics and Science Study), in 2003 (as reported by Martin, *at al.*, 2004). In each of them the main position effect can be related to speed effect that is quite different from fatigue effect because it occurs when students do not have enough testing time (Martin, *et al.,* 2004). In any case, item position effects seem to be dependent also on the measured competence, because findings for verbal tests yield different results (e.g., Leary & Dorans, 1985; Zwick, 1991).

In any case, aside from those empirical evidences, the study of item-position effect is still topical first at all from a methodological and statistical point of view. In fact, in order to assess students' achievement, the most frequently used tool is the Rasch model, according to which the probability of a correct answer is governed by students' relative ability, i.e. student intrinsic ability ($\beta_n$) compared to item difficulty ($\delta_i$). Moreover, according to the model, item difficulty depends only on item content and any other item characteristics (that concur in altering its difficulty) works as unexpected (i.e. not explicitly hypothesized) disturbing factor. Therefore, it must be removed in order to avoid biases in both item and person parameter. In fact, even if «(...) any item position can be balanced over all the test booklets, as a consequence of which

the averaged ability parameter estimation becomes unbiased (...) some large-scale assessments provide additional feedback to every individual examinee, as well as sometimes to the individual class or school; in this case, when several test booklets have been used for instance to limit cheating, any item-position effect would invalidate the individuals' test results» (Hohensinn, et al., 2008, 392).

***Relationship between item-position effects and Rasch model***

The Rasch model (1960; 1980) is the most frequently used tool "to measure" students' abilities. It hypothesizes that subject's answer to an item depends only on his/her relative ability, i.e. student's *intrinsic* ability compared to item's difficulty.

The Rasch model hypothesizes that item difficulty has be determined only by its content and no other factor can concur in explaining it. In fact, if this happens, item functionality hypothesized by the Rasch model can change, causing bias in both item and person parameter (Embretson & Reise, 2000). And, item-position within a test can work as "disturbing" factors (as previously pointed out by Kubinger, 2009).

According to the literature, there are two main possible kinds of item-position effects, i.e. *learning* and *fatigue* effect.

The first one can take place when students become familiar with the test material and the kind of tasks while working on the test. As consequences, they become more acquainted with test-taking strategies which can improve the test score (Rogers & Yang, 1996). Obviously, this effect gets stronger when there are some items that prompt the correct answer for following items. Generally, the second case implies a violation of local independence (one of the theoretical assumptions of the Rasch model – Hambleton & Swaminathan, 1985), according to which the probability that subject *n* gives a correct answer to the item *i* ($\Pr\{x_{ni}=1\}$) is independent of the probability of correctly answering each other items embedded in the same achievement test ([2]).

$$P\left(X_n \mid \beta_n, \delta_1, \delta_2, \delta_3 ..., \delta_i\right) = \prod_{i=1}^{k} P\left(X_{ni} \mid \beta_n, \delta_i\right) \qquad [2]$$

Nevertheless, also the *fatigue* effect can be considered as a violation of the model because, when it occurs, students do not give the correct answer not because they do not have the necessary quantity of ability to do it, but just because they are too tired. In fact, a fatigue effect occurs when an item at the beginning of a test is less difficult than the same item administrated at the end of the same test because students' attention and reactivity usually decrease over time.

In any case, item position can be considered also in a different manner. As suggested by literature, item difficulty can be expressed as function of several typical "factors". For the primary school, for example they are «(...) the size of the largest and smallest number occurring in the problem, the form of the equation to be solved, the number of arithmetic operations to be performed (...)», and so on (Suppes, 1968). According to this perspective, item difficulty can be explained as function of cognitive processes activated by each student in order to solve it. About that, Fischer claimed that «(...) measurement models which describe the difficulty of each learning unit, of each task or "item", as a function of the basic operations involved, can

be devised and empirically tested in a relatively simple way» (Fischer, 1973, 360).

***Research Hypotheses***

In the light of what said above, we tested the following hypotheses:

**HP1**: Item functionality changes depending on its position within the test: *ceteris paribus*, item parameter as well as observed probability for equivalent groups of students (i.e., students with the same ability level) changes depending on item position within the test.

**HP2**: Overall test functionality changes if its structure (i.e., item administration order) changes. Implications of this can be, for example, that differences in difficulty parameter estimation that imply biased ability parameter estimations can occur, or that, starting from two items batteries, composed by the same items, simply administered in reverse order, scale construction process can lead to different results (e.g. to the identification of different misfitting items) and, as consequence, to different scales.

## Methodology

***Data***

The analyses presented in this study was carried out on secondary data, collected, for different purposes, by the Italian National Institute for the Evaluation of Educational System, by administering to pupils attending the $5^{th}$ grade levels of primary school two achievement tests, named A and B respectively, aimed at measuring mathematical ability.

Both A and B were aimed at measuring students' ability in four main domains, i.e. numeracy, geometry, data and prevision (i.e., elements of Statistics), and a fourth dimension named "Relations and functions", aimed at: verifying students' ability in classifying objects, figures, and so on; doing and recognizing equivalences and orderings; understanding and applying to real cases both direct and inverse proportionality concepts; comparing situations, figures or sequences in order to recognize similarities/similes, and so on; as specified by INVALSI in its "Quadri di Riferimento (Qdr)" (*Reference Frameworks*) for primary school (INVALSI, 2010).

The first test (A) consisted of 52 items. In our design, 273 pupils are randomly assigned the form to be administered by means of spiralling process. Test A was administered in two forms, named A1 and A2 respectively, by means of spiralling process. A *spiraling* process is one procedure that can be used to randomly assign forms using this design. When the booklets are handed out, the first examinee receives Form A1, the second examinee Form A2, the third examinee Form A1, and so on. This spiraling process typically leads to comparable, *randomly equivalent* groups taking Form A1 and Form A2. For this reason, «when using this design, the difference between group level performance on the two forms is taken as a direct indication of the difference in difficulty between the forms». (Kolen & Brennan, 2004, 13). Both forms were composed by the same items, administered in reverse order. By means of this design, on average, an half of the sample took A1 and the other half took A2.

The second test (B) consisted of 50 items. It was administered to 259 pupils. By the means of spiralling process, half of students are randomly assigned the form B1 and half of them the form B2, both ones

composed by the same items, administered in reverse order, as in the previous case.

In both cases, the percentage of male and female students was equally distributed as well as other socio-demographic variables, such as, for example, parental education and occupation. Each participant was randomly assigned a test form. The test administration was not speeded and participants were allowed up to 75 minutes to complete the test form. Nearly all participants completed the test in the allotted time.

***Methods***

In order to examine item-position effects, a lot of different methods have been proposed, for example, ones based on mean comparison (Marso, 1970) or correlations such as confirmatory factor analysis (Schweizer, Schreiner, & Gold, 2009), as well as some models within the Item Response Theory such as the Linear Logistic Test Model (LLTM by Fischer, 1973), an extension of the Rasch model, usually used to model item attributes like presented position, and test their impact on item difficulty (Kubinger K. D., 2008; 2009) as well as to model which elementary cognitive components are necessary for solving a certain item (Embretson & Daniel, 2008; Messick, 1995).

In the present study, the methodological strategy, carried out in the framework of Item Response Theory, was divided in three phases.

*Phase I.* Starting from the complete items batteries, for each version of the same test, misfitting items were individuated. To do this, a fit analysis based on Chi-Square Test was carried out by using software RUMM2030. It also provides some preliminary information that can be used to have an overview on items battery functionality: 1. Cronbach α (that is a measure of internal consistency of items' battery); and, 2. Person Separation Index (PSI, that is the ratio of the unbiased estimate of the sample standard deviation to the root mean square measurement error of the sample, i.e. it quantifies "reliability" and measurement precision - Andrich, 1982).

*Phase II*. Both items and test functionality was analyzed. For the first one, item parameters were estimated in order to understand if and how some changes in items' position could affect their difficulties. For test functionality, a comparison between item location orders along the latent trait both for A1 relative to A2 as well as for B1 relative to B2 was carried out. Spearman rank-order correlation (ρ), a non-parametric version of the Pearson product-moment correlation, was used to measure the strength of association between two ranked variables: the closer the value is to 1, the better the correlation, and thus more similar scales are between themselves from a psychometrical point of view.

*Phase III*. The comparison between item location orders carried out on the basis of estimated ability could led to misleading results, in particular with reference to different ability (sub-)groups of students (e.g., unexpected compensations between the best and the worst performances may occur and this circumstance could lead to - just apparent - similar results in item parameter estimations). In order to disclose these differences, graphical inspection of Items Characteristics' Curves (ICCs) that shows item behavior in each (sub-) group of students clustered by ability. As consequence, by the means of comparison between these two curves, we could say that, *ceteris paribus*, if item behavior changes for students with the same ability when just its position within the test changes, thus item position affects its functionality.

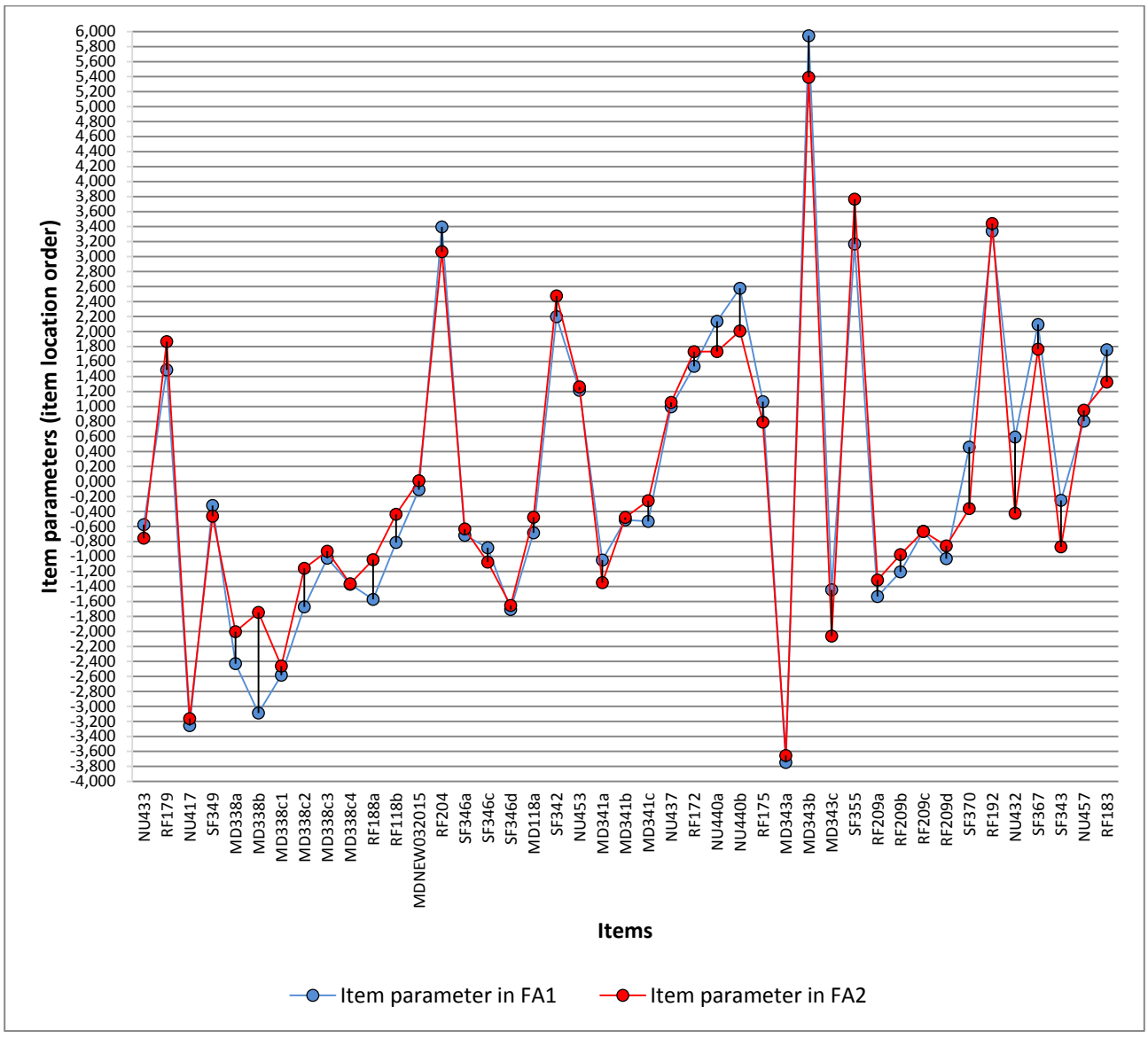
## Results

The analysis carried out to point out misfitting items gave back different results both for A1 relative to A2 as well as for B1 relative to B2. This first empirical result, in absence of further variables that can help in explaining it, seemed suggest that item position within the test might cause these differences.

Nevertheless, in order to make possible a comparison item by item, we selected a subset of items that guaranteed adequate fit level and formed a Rasch scale for both A1 and A2 as well as for both B1 and B2. All the other items were excluded.

In the first case (test A), item parameter distributions, estimated just on common items, (figure 1) showed a little fatigue effect for some items (e.g., MD338b that was in position 24 in A1, and in position 5 in A2; NU432 in position 47 in A1 and position 6 in A2; SF355 that was in position 6 in A1, and 20 in A2; and so on).

**Figure 1 - Item difficulties estimated by the Rasch model for each item pair (FA1 – FA2).**




*Source*: our elaboration

The magnitude of these differences was not so large: by comparing item location order in A1 and A2, it is evident that they were substantially the same both in A1 and A2 (Table 1), as also confirmed by Spearman rank-order correlation test [ρ (98)= 1,00, p < .0005] (Table 2).

**Table 1 - Comparison between item location order along the latent trait in booklet A1 and booklet A2.**

| **Descriptives** | **FA1** | **FA2** |
|---|---|---|
| Mean | 0,00 | 0,00 |
| Standard deviation | 2,02 | 1,90 |
| Variance | 4,07 | 3,61 |
| Minumun | -3,75 | -3,65 |
| Maximum | 5,95 | 5,39 |

*Source*: our elaboration

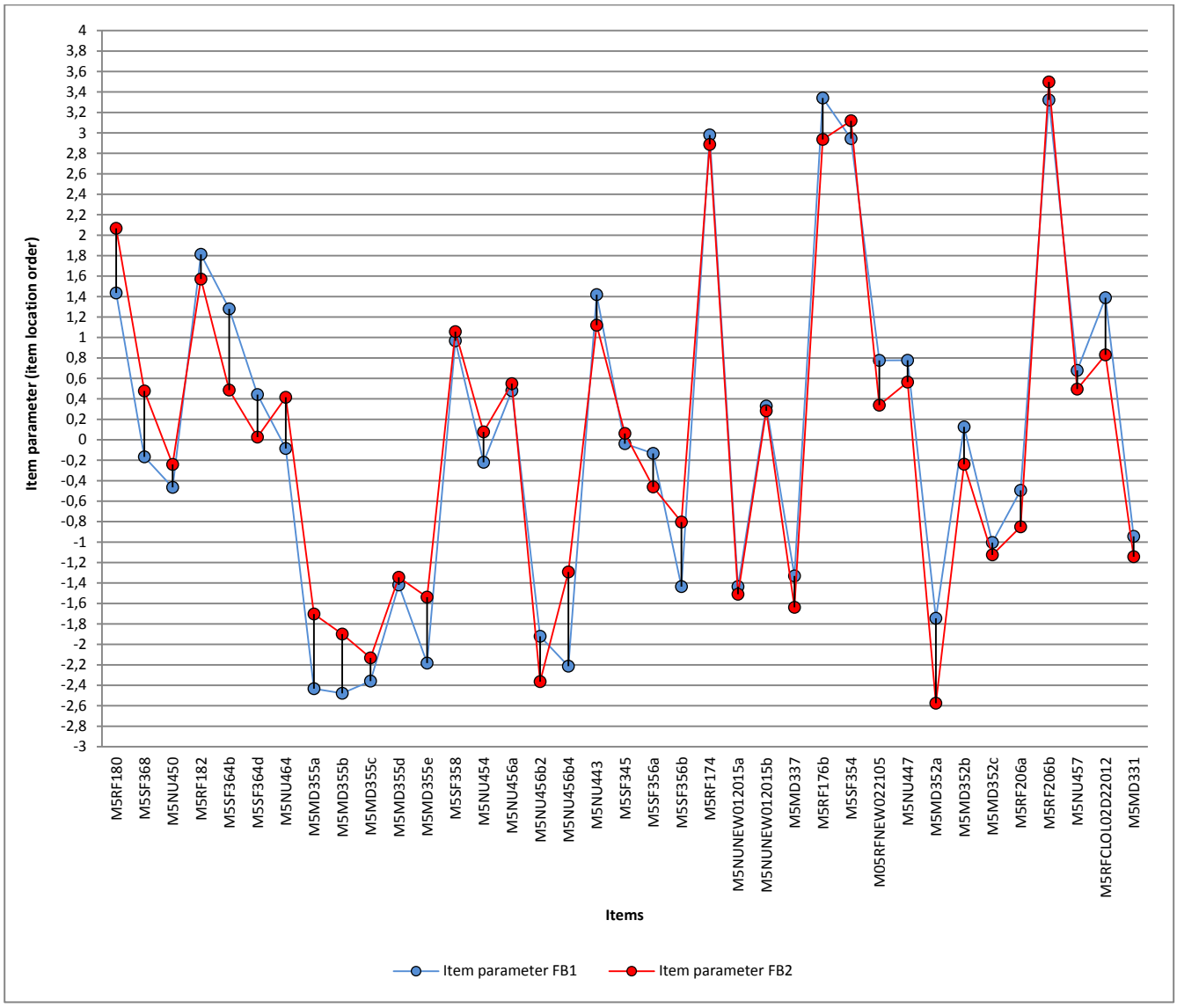
**Table 2 - Spearman rank-order correlation test.**

| | **Item difficulty (A1)** | **Item difficulty (A2)** |
|---|---|---|
| Correlation coefficient | 1 | 1,000** |
| Sign. (two-fold) | . | . |
| N | 43 | 43 |
| Correlation coefficient | 1,000** | 1 |
| Sign. (two-fold) | . | . |
| N | 43 | 43 |

**. Statistically significant correlation at 0,01 (two-fold test).
*Source*: our elaboration

The analysis carried out taking into account B1 and B2 showed very similar results, although differences in item parameters estimation were deeper than ones observed in the previous case.

**Figure 2 - Item difficulties estimated by the Rasch model for each item pair (FB1 – FB2).**




Notwithstanding this difference, items location order along the latent trait was substantially the same both in B1 and B2 (Table 3), as also confirmed by Spearman rank-order correlation test [$\rho(98)= 1{,}00$, $p<.0005$] (Table 4).

**Table 3 - Comparison between item location order along the latent trait in booklet A1 and booklet A2.**

| **Descriptives** | **B1** | **B2** |
|---|---|---|
| Mean | 0,00 | 0,00 |
| Standard deviation | 1,65 | 1,57 |
| Variance | 2,73 | 2,47 |
| Minumun | -2,48 | -2,58 |
| Maximum | 3,34 | 3,50 |

*Source*: our elaboration

**Table 3 - Spearman rank-order correlation test.**

| | Item difficulty (B1) | Item difficulty (B2) |
|---|---|---|
| Regression coefficient | 1,000 | 1,000** |
| Sign. (two-fold) | | ,000 |
| N | 37 | 37 |
| Regression coefficient | 1,000** | 1,000 |
| Sign. (two-fold) | ,000 | |
| N | 37 | 37 |

**. Statistically significant correlation at 0,01 (two-fold).
*Source*: our elaboration

### *Graphical inspection of Item Characteristics Curves (ICCs)*

The analysis carried out in order to pick out information about both items and test functionality did not disclose any significant differences between A1 and A2, and between B1 and to B2. In addition, results shown in the previous paragraph were in line with what argued by the most recent literature (e.g., Hohensinn & Kubinger, 2011). In any case, they were not really surprising since our testing items were the same both in A1 and A2 as well as in B1 and B2.

Despite this, fit analysis carried out in the first phase of our study led to picking out different misfitting items for A1 and A2 as well as for B1 and B2. Therefore, despite of results showed in the previous paragraph, this empirical evidence seemed to suggest that, *ceteris paribus*, items administration order could affect items behavior and, as consequence, alter scale construction process. (For example, as said above, in order to compare item functionality in the present study, we just considered items that guaranteed adequate fit level both for A1 and for A2 as well as both for B1 and for B2).

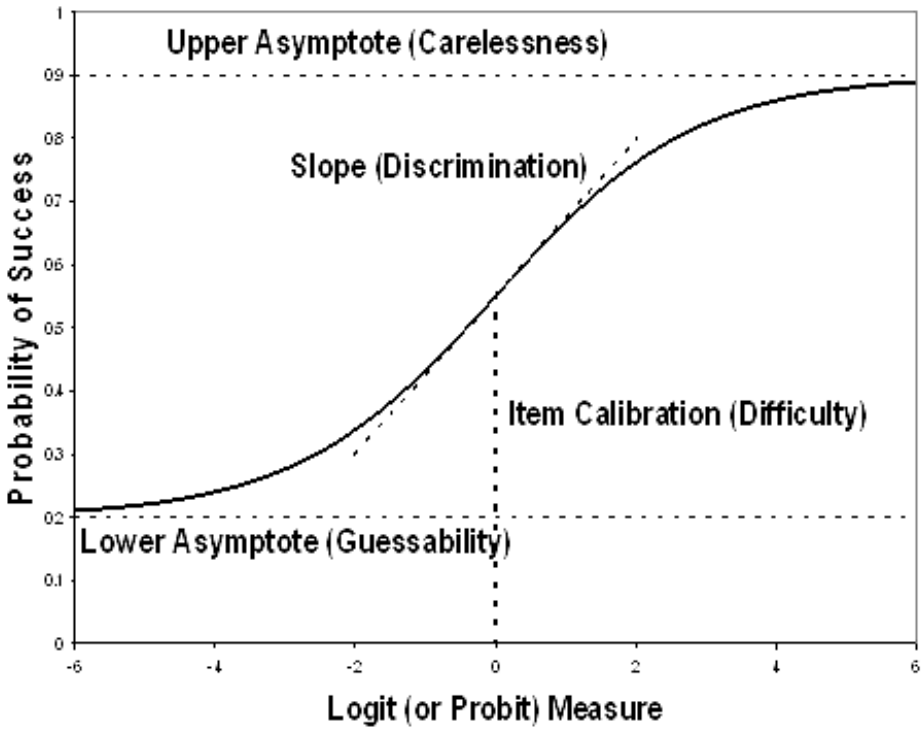
To verify this conjecture, a comparison between items' characteristic curves (ICCs) was carried out. Differently from other measures that statistically "describe" an item (fit measures as well as its difficulty, and so on), the comparison of ICCs gave deeper information about its functionality, first at all, because ICC was able to highlight difference in item behavior for different ability sub-groups of students.

In fact, as said above, the probability of correct response depends on students' relative ability, i.e. their intrinsic ability compared to items difficulty. This means that the probability of success is near zero at the lowest levels of ability, and it approaches 1 when ability increases (Figure 3). This logistic curve (known as the *Item Characteristic Curve* within Item Response Theory framework) describes the relationship between the probability of correct response and the ability scale.

RUMM2030, as well as other similar software, plots ICC for each item in a test, and describes its functionality for each ability level. This is very important because, differently from 2PL and 3PL IRT models that provide *ad hoc* measures for both discrimination and guessing effect (very important aspect in order to study item functionality), within 1PL model, the graphical inspection of ICC allows to study also item discrimination as well as guessing. Item discrimination refers to the possibility of correctly distinguishing subjects depending on the real quantity of ability owned by each of them. This item feature is very important because from it results the possibility of correctly scaling subjects along the latent trait. From a techincal point of view, in the context of 1PL model, item discrimination can be evaluated by means of ICC's slope. The ICC is very useful also to quantify guessing effect, i.e. the possibility of unearned success by lucky guessing.

Obviously, both discrimination and guessing are very important aspects of item behavior with relevant implications on overall scale functionality.

**Figure 3 - A typical item characteristic curve (ICC).**



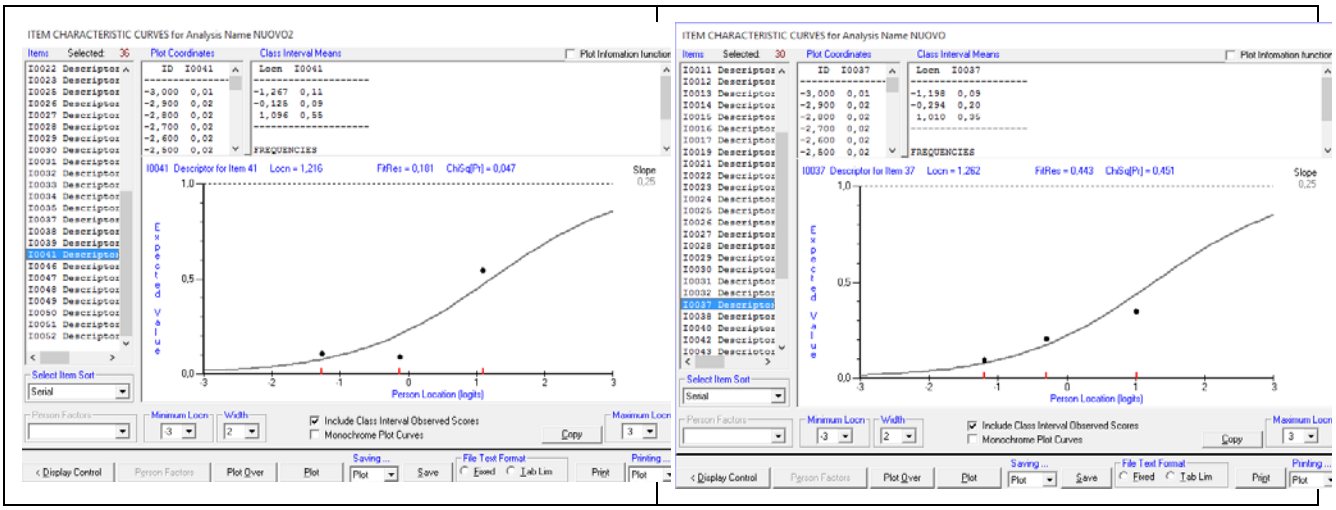

*Source:* http://www.rasch.org/rmt/rmt181b.htm

RUMM2030 plots ICCs in a very useful manner. Black points within the graphs represent sub-groups of students clustered as function of their estimated ability level. So, for each of them, the probability of correct answer can be clearly observed. Corresponding points on the theoretical curve are the estimated probability of correctly answer that item for each ability level. The distance between estimated and observed probabilities can be used as a measure of how well data fit the model and it describes observed item functionality relative to model expectations.

In this direction a clear example was given by the item M5NU453 that appears in the middle part both of A1 and of A2. Although no strong difference in item difficulty could be revealed ($\delta A1 = 1.216$; $\delta A2 = 1.262$, in the observed empirical range [-2.5; +3.5] on the latent trait), item functionality in A1 and A2 was clearly different (In this sense, doubtless, the graphical inspection of ICCs used to explore differences in item functionality gives back a significantly more precise information then the mere comparison between item parameters).

In A1, data did not very well fit model's expectation, especially for the middle ability students group. On the contrary, in A2, although model clearly overestimates the probability of a correct answer for the more skilled pupils, it better predicted the overall item functionality, especially for the other ability students' levels.

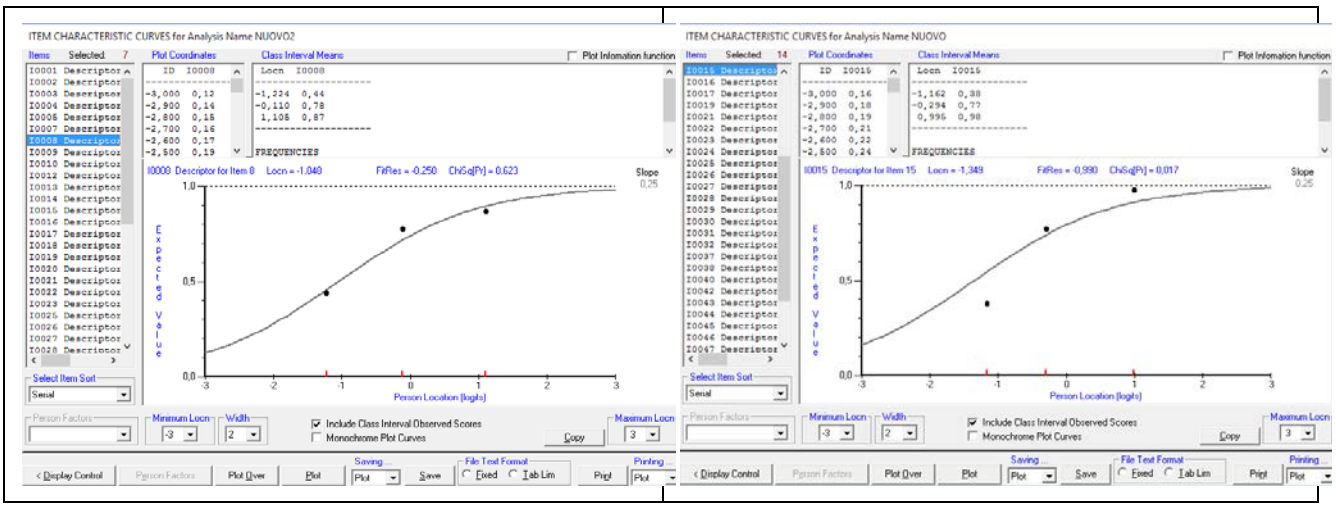
**Table 4 - M5NU453's ICC in A1 (graph on the left) and A2 (graph on the right)**




*Source*: our elaboration

Interesting differences in item functionality could be clearly disclosed also for the item M5MD341a, administered in position 15 in A1 and in position 8 in A2.

**Table 5 - M5MD341a's ICC in A1 (graph on the left) and A2 (graph on the right)**




*Source*: our elaboration

Also in this case, although no strong difference in item difficulty could be revealed ($\delta_{A1}$ = -1.048 and $\delta_{A2}$ = -1.349), circumstance that concurred in explaining little guessing effect in both cases, item functionality in A1 and A2 is quite different. Answers given by students to M5MD341 were more coherent to model's assumptions in the context of A1 test booklet relative to A2, as also confirmed by fit residual and Chi-Square(Pr) ($FitRes_{A1}$ = -0.250 and $\chi^2$ $(Pr)_{A1}$ = 0.623 versus $FitRes_{A2}$ = -0.990 and $\chi^2(Pr)_{A2}$ = 0.017).

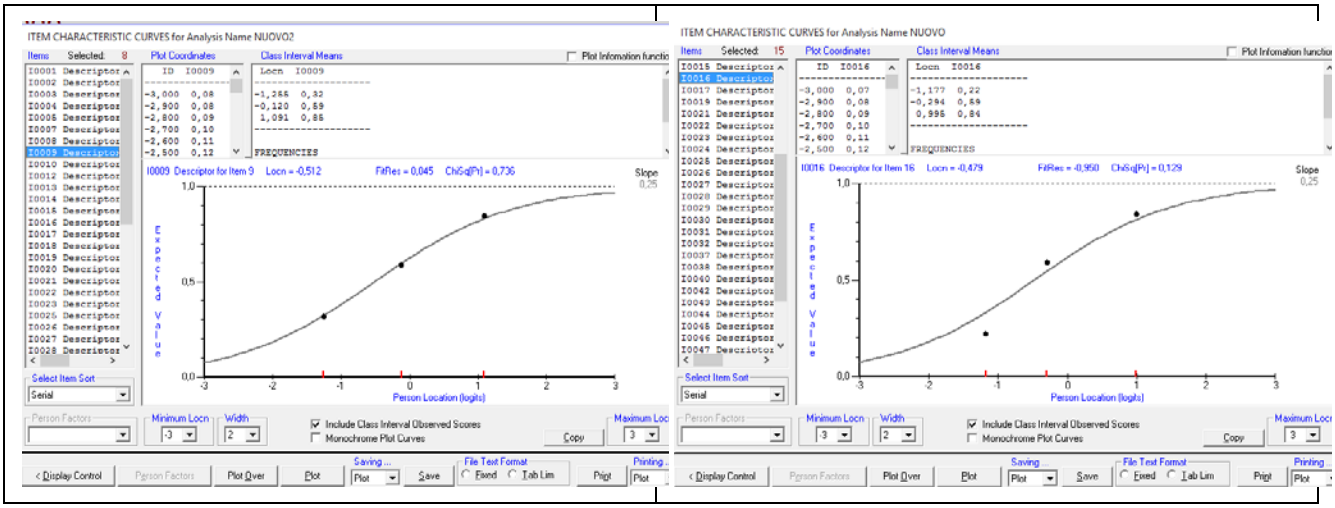
In A2, M5MD341a represented a case of over discrimination, i.e. model overestimated the probability of a correct answer for low ability students' group, and vice versa for the more skilled pupils. On the contrary, in A1, observed answers were almost completely coherent to model's expectations. An unsatisfying fit level was highlighted also by Chi-Square Probability, equal to 0.623 in A1 and 0.017 in A2 (and thus less than 0.05). Just a little distance between estimated and observed probability could be disclosed for the middle group, in A1 as well as in A2.

Over discrimation is usually caused by a violation of *local independence* that means that, after taking into

account examinee ability, his/her responses to the items are statistically independent. «(...) The local independence assumption implies that there are no dependencies among items other than those that are attributable to latent ability. One example where local independence likely would not hold is when tests are composed of sets of items that are based in common stimuli, such as reading passages or chart. In this case, local independence probably would be violated because items associated with one stimulus are likely to be more related to one another than to items associates with another stimulus. (...) Although the IRT unidimensionality and local independence assumptions might not hold strictly, they might hold closely enough for IRT to be used advantageously in many practical situations» (Kolen & Brennan, 2004, p. 157). The cited passage referred to local independence as violation of *response* independence (Marais & Andrich, 2008): it typically occurs when the probability of correctly answering an item depends on the presence of another item in the test, for example, because, this item (implicitly or explicitly) focalized students' attention on one or more things/aspects/etc. that leads pupils to reach the correct solution, although they are not good enough to correctly answer to that item.

Similarly to M5MD341a, also the subsequent item M5MD341b, referred to the same stimulus, showed some difference in its functionality. In A1, data are perfectly coherent to model assumptions (FitRes = 0.045, Chi-Square (Pr) = 0.736). Instead, in A2, the model overestimates the probability of correct answer for less skilled students.

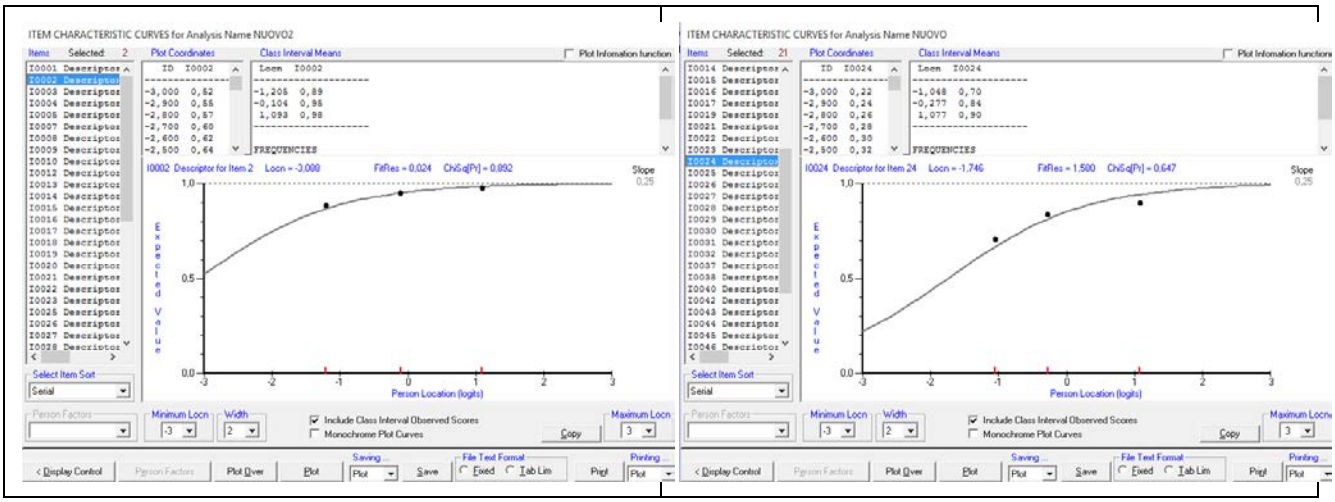
**Table 6 - M5MD341b's ICC in A1 (graph on the left) and A2 (graph on the right)**




*Source*: our elaboration

Another interesting case was represented by the item M5MD338b administered after another item that referred to the same stimulus.

**Table 7 - M5MD338b's ICC in A1 (graph on the left) and A2 (graph on the right)**




*Source*: our elaboration

In this case, a significant difference in item difficulty could be disclosed ($\delta_{A1}$ = -3.088 versus $\delta_{A1}$ = -1.076) that can be easily explained taking into account that this item was administered in position 24 in A1 and in position 2 in A2. In addition, a relevant difference can be disclosed in the theoretical curve shape. First at all, it meant that M5MD338b discrimination was lower in A1 than in A2. Item discrimination refers to item "ability" in distinguishing between students depending on their ability levels, and, thus, scaling them on the latent trait correctly. In this sense, «item discrimination indicates the extent to which success on an item corresponds to success on the whole test. Since all items in a test are intended to cooperate to generate an overall test score, any item with negative or zero discrimination undermines the test. Positive item discrimination is generally productive, unless it is so high that the item merely repeats the information provided by other items on the test» (Kelley, Ebel, & Linacre, 2002).

In addition, a strong guessing effect could be disclosed both in A1 and in A2. In first case, the theoretical curve intercepted y axis around 50%. This meant that also the worst student had a probability of correctly answering this item higher than 50%, although it was administered in position 24. In this sense, hypothesizing that M5MD338b functionality was affected by learning and fatigue effect seemed plausible. In this sense, a confirmation seemed resulted by the minor guessing effect observed for the same item located at position 2 in A2. In this case, the theoretical curve intercepted y axis at around 20%. As in the previous case, this meant that the worst students had a probability of correctly answering this item higher than 20%, i.e. 30% less than students that faced the same item located in position 24 in A2.

## Discussion and conclusive remarks

The diffusion of the Rasch model in Psychometrics is due to a lot of different reasons, such as its intrinsic easiness, the data quality (it produces quasi-metric ability measure), and its statistical properties.

In order to guarantee that model's properties hold, data must fit the model, i.e. its theoretical assumptions (Hambleton, Swaminathan, 1985): 1. *Unidimensionality* (both items and subjects are scaled along the same latent trait – generally, an ability or a behavior – depending, respectively, on $\beta_n$ and $\delta_i$); 2. *Monotonicity* (the probability of a correct answer is a monotone function of ability, i.e. students with a higher ability level has a higher probability of giving a correct answer); and, 3. *Local independence* (the $Pr\{x_{ni}=1\}$ is independent of the probability of correctly answering to each other items embedded in the same achievement test).

Local independence «(…) means that, after taking into account examinee ability, examinee responses to the items are statistically independent. Under local independence, the probability that examinees of ability β correctly answer both item 1 and item 2 equals the product of the probability of correctly answering item 1 and the probability of correctly answering item 2. For example, if examinees of ability β=1.5 have a .5 probability of answering item 1 correctly and .6 probability of answering item 2 correctly, for such examinees the probability of correctly answering both items correctly under local independence is .30=.50(.60) (Wright & Linacre, 1994).

The local independence assumption implies that there are no dependencies among items other than those that are attributable to latent ability. One example where local independence likely would not hold is when test are composed of sets of items that are based in common stimulus, such as reading passages or chart. In this case, local independence probably would be violated because items associated with one stimulus are likely to be more related to one another than to items associates with another stimulus, simply because, for example, they can contain some (explicit or implicit) suggestions to solve next items (related to the same stimulus). Nevertheless, the same mechanism could work also in different circumstances, e.g. when the same kind of graph is used to write two different items within the same test, and so on.

In this sense, the "concept" of item position does not just refer to its (single) position within the test but it more correctly refers to relative item position, i.e. its position compared to items ordering within the entire test. In this sense, some students might have an advantage working on a certain item at a certain position, while the other might be handicapped. This could happen because if there are different sequences of item presentation, different solving strategies (i.e. based on different cognitive processes) might be activated by students. In this sense, item position can affect its functionality as well as the entire test functionality because different item administration order leads students to develop different solving strategies and thus to activate different cognitive processes, although achievement test are composed by the same items.

Results presented in the previous paragraph were just an example of how the graphical inspection of ICC works. This kind of analysis guaranteed finer results than the traditional strategy based on the comparison of item parameter estimation because it was able to reveal significant differences in item functionality also when no significant differences in its difficulty could be disclosed.

In fact, the graphical inspection of ICC confirmed that item position affected item functionality and that this

circumstance had relevant effect on overall items battery. Examining the relationship between items position and their functionality deserve considerable emphasis, because problems with the ascertainment of this relation can have serious consequences. If item estimation fails, then all findings can lead to erroneous conclusions. These results, in fact, (at least) concurred in explaining why, starting from two batteries composed by the same items, misfitting items were different in A1 relative to A2 as well as in B1 relative to B2.

Nevertheless, more investigations are surely needed, first at all because, in order to really understand the mechanism through which items position could concur in modifying their functionality, a deeper analysis carried out item by item, for example also by the means of direct interviews to pupils, might show if a relationship between items administration order and solving strategy actually exist.

In order to really explain results presented above, this step is absolutely important. For each item, students activate specific cognitive processes (Fischer, 1973). According to the hypothesis that we would like to test in the future, the activation of some processes might determine the activation or the non activation of next processes, and thus concurs in specifying the overall solving strategy carried out by students to face all items embedded in the achievement test. This topic seems absolutely interesting because test functionality depends on items behavior. Therefore, understanding the mechanism through which relative item position (in the meaning specified above) affects items functionality could be used in order to better organize items within an achievement test and thus in order to preserve its wished psychometrical properties.

References

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index and the Guttman scale response pattern. *Education Research and Perspectives, 9*, 95-104.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Qurterly, 50*(3), 328-344.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah, New Jersey: Lawrence Erlbaum associates Publishers.

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

Fischer, G. H., & Pendl, P. (1980). Individualized Testing on the Basis of Dichotomous Rasch Model. In L. J. van der Kamp, W. F. Langerak, & D. N. de Gruijter, *Psychometrics for Educational Debates* (pp. 171-187). Chichester, England: John Wiley Et Sons.

Flaugher, R. L., Melton, R. S., & Myers, C. T. (1968). Item rearrangement under typical test conditions. *Educational and Psychological Measurement, 28*(3), 813-824.

Giampaglia, G., & Guasco, B. (2011). La rilevazione degli apprendimenti nelle scuole italiane: un'analisi dei dati invalsi . *Polena*.

Green, K., & Smith, R. M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics, 12*, 369-381.

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly, 50*, 379-390.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston: Kluwer-Nijho.

Hoehnsinn, C., & Kubinger, K. D. (2009). On Varying Item Difficulty by Changing the Response Format for a Mathematical Competence Test. *Austrian Journal of Statistics, 38*(4), 231-239.

Hoehnsinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly, 50*(3), 391-402.

Hohensinn, C., & Kubinger, K. D. (2009). On varying item difficulty by changing the response format for a mathematical competence test. *Austrian Journal of Statistics, 38*(4), 231-239.

Hohensinn, C., & Kubinger, K. D. (2011). Applying Item Response Theory Methods to Examine the Impact of Different Response Formats. *Educational and Psychological Measurement, 71*(4), 732–746.

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly, 50*(3), 391-402.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysis item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation: An International Journal on Theory and Practice, 17*(6), 497 - 509.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and Practices.* New

York: Springer (2nd ed).

Kubinger, K. (2005). Psychological Test Calibration Using the Rasch Model. Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5*(4), 377-394.

Kubinger, K. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly, 50*(3), 311-327.

Kubinger, K. (2011). Applications of the Linear Logistic Test Model in Psychometric Research. *Educational and Psychological Measurement, 1*(71), 732-746.

Kubinger, K. D. (2005). Psychological Test Calibration Using the Rasch Model - Some Critical Suggestions on Traditional Approach. *International Journal of Testing, 5*(4), 377-394.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructiong tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly, 50*(3), 311-327.

Kubinger, K. D. (2009). Applications of the Linear Logistic Test Model in Psychometric Research. *Educational and Psychological Measurement*.

Kubinger, K. D., & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on different item response formats - An experiment in fundamental research on psychological assessment. *Psychology Science, 49*(4), 361-374.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*, 387-413.

Mair, P., & Hatzinger, R. (2007). Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software, 20*(9), 1-20.

Marais, I., & Andrich, D. (2008). Formalising dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(3), 200-2015.

Marso, R. N. (1970). Test item arrangement, testing time, and performance. *Journal of Educational Measurement, 7*, 113-118.

Martin, M. O., Mullis, I. V., & Chrostowski, S. J. (2004). Item analysis and review. In M. O. Martin, I. V. Mullis, & S. J. Chrostowski, *TIMMSS 2003 technical report* (pp. 224-251). Chestnut, MA: TIMSS & PIRLS International Study Center, Boston College.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 741-749.

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika, 15*(3), 291-315.

Mullis, I. V., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report.* Boston: Chestnut Hill, MA: Boston College.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: The university of Chicago Press.

Rasch, G. (1961). On general laws and meaning of measurement in psychology. *Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Theory of Probability* (pp. 321-333). Berkeley:

University of California Press.

Rasch, G. (1977). On Specific Objectivity: An Attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58-94.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests (Reprint).* Chicago: The University of Chicago Press.

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology, 103*(1).

Rogers, W. T., & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment, 12*, 247-259.

Sax, G., & Cromack, T. R. (1966). The effects of various forms of item arrangements on test perormance. *Journal of Educational Measurement, 3*(4), 309-311.

Schweizer, K., Schreiner, M., & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: A two-dimensional model of APM. *Psychology Science Quarterly, 51*, 47-64.

Suppes, P. (1968). *Computer-Assisted Instruction.* Standford's 1965-66 Arithmetic Program.

Wu, M. (2010). Measurement, Sampling, and Equating Errors in Large-Scale Assessments. *Educational Measurement: Issues and Practice, 29*(4), 15-27.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 55*, 387-413.