Psicothema

Psicothema 2016, Vol. 28, No. 3, 346-352
doi: 10.7334/psicothema2015.287

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

A multistage adaptive test of fluid intelligence

Manuel Martín-Fernández, Vicente Ponsoda, Julio Olea, Pei-Chun Shih and Javier Revuelta
Universidad Autónoma de Madrid

Abstract

Background: Multistage adaptive testing has recently emerged as an alternative to the computerized adaptive test. The current study details a new multistage test to assess fluid intelligence. **Method**: An item pool of progressive matrices with constructed response format was developed, and divided into six subtests. The subtests were applied to a sample of 724 college students and their psychometric properties were studied (i.e., reliability, dimensionality and validity evidence). The item pool was calibrated under the graded response model, and two multistage structures were developed, based on the automatic test assembly principles. Finally, the test information provided by each structure was compared in order to select the most appropriate one. **Results**: The item pool showed adequate psychometric properties. From the two compared multistage structures, the simplest structure (i.e., routing test and two modules in the next stages) were more informative across the latent trait continuum and were therefore kept. **Discussion:** Taken together, the results of the two studies support the application of the FIMT (Fluid Intelligence Multistage Test), a multistage test to assess fluid intelligence accurately and innovatively.

Keywords: multistage testing, automated test assembly, fluid intelligence.

Resumen

Un test adaptativo multietapa de inteligencia fluida. **Antecedentes:** los test adaptativos multietapa han emergido recientemente como una alternativa a los test adaptativos informatizados. Se presenta en este estudio un test multietapa para evaluar la inteligencia fluida. **Método:** se desarrolló un banco de ítems de matrices progresivas con formato de respuesta construida que posteriormente fue dividido en seis subtests. Los ítems se administraron a un total de 724 estudiantes universitarios. Se estudiaron las propiedades psicométricas de los subtests (fiabilidad, dimensionalidad, evidencias de validez) y se calibró el banco con el modelo de respuesta graduada. Se construyeron después dos estructuras multietapa a través del ensamblaje automático de tests y se comparó la información proporcionada por cada una de ellas. **Resultados:** los ítems mostraron unas propiedades psicométricas adecuadas. De las dos estructuras puestas a prueba, se conservó finalmente la estructura sencilla, pues resultó más informativa. **Discusión:** los resultados de estos dos estudios avalan el empleo del FIMT, una herramienta que emplea este formato para evaluar de forma innovadora y precisa la inteligencia fluida.

Palabras clave: test multietapa, ensamblaje automático de test, inteligencia fluida.

Multistage adaptive testing has reemerged in the last few years as an alternative to computerized adaptive tests (CAT). Unlike CAT, multistage tests administer a predetermined set of items (i.e., modules) to respondents at each adaptation point. Therefore, adaptation takes place between modules and not between single items. However, the main difference between a multistage test and a CAT is that all possible test forms can easily be constructed before administration of the items begins. Thus, a strong control can be assumed over the attributes of the items (such as item content or item difficulty) comprising each module (Hendrickson, 2007; Yan, von Davier, & Lewis, 2014). In multistage adaptive tests, examinees go through several stages; they take a first stage with a module of moderate difficulty, or highly informative for medium latent trait levels, often called the routing test, and, depending on their performance, they are sent to one of the modules of the second stage, more adjusted to the level of each respondent in the evaluated latent trait. Therefore, once a module is completed, adaptation takes place until the last stage is reached.

Different multistage structures can be assembled. A multistage structure with a single module in the first stage and three modules in the second stage is called a 1-3 structure. Thus, a structure with three stages with one module in the first stage and two modules in the second and third stages is a 1-2-2 structure. Furthermore, as the number of stages and the modules per structure increase, so do the complexity and the adaptability of the structure. Nonetheless, Luecht and Nungester (1998) pointed out the relevance of finding a balance between the complexity and the adaptability of a multistage test. Adding too many stages and modules could not lead to optimal structures. In fact, it seems that structures with a low number of modules per stage perform as well as structures with the same number of stages and more modules per stage (Wang, Fluegge, & Luecht, 2012).

The procedure usually followed to assemble a multistage test is the Automated Test Assembly (ATA; Diao & Van der Linden, 2011). ATA converts the test construction into a linear optimization problem. An objective function (i.e., the test information) has to be maximized or minimized with respect to some variables (i.e., items to be selected) fulfilling several constraints (i.e., the item content of the modules, item difficulty, or the test length). Thus, we

Received: November 6, 2015 • Accepted: May 10, 2016
Corresponding author: Manuel Martín-Fernández
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
e-mail: manuel.martinf@inv.uam.es

could determine which items should compose the modules of each stage of the test to reach the optimal solution.

In this study, ATA is used to assemble a multistage test with a pool of progressive matrix items. Like the Raven test, these items attempt to measure the fluid intelligence of the examinees. Progressive matrix tasks are a key component in the evaluation the *g*-factor or general mental ability (Snow, Kyllonen, & Marshalek, 1984). Abstract, numerical, and spatial reasoning are some of the usual abilities more closely related to matrix tests (Colom, Escorial, Shih, & Privado, 2007). Lower relations can be also found between the matrix tests and other measures of crystallized intelligence, such as verbal reasoning or vocabulary tasks (Colom, Abad, Quiroga, Shih, & Flores-Mendoza, 2008). Furthermore, the relation between fluid intelligence and working memory was widely studied in the academic literature during the last two decades (Ackerman, Beier, & Boyle, 2005).

The purpose of this research is to assemble a new and fully operational fluid intelligence multistage test. This test presents two peculiarities: (a) it is based on a short and heavily constrained item pool, and (b) it is meant to be applied in evaluation contexts for high-ability level examinees. The strategies followed to deal with such an item pool and to maximize test information for high latent trait values are detailed in the following studies.

Study 1

The first study focuses on the development and the psychometric properties of the initial item pool, such as reliability, dimensionality and validity evidence based on relations with other related variables. An IRT model is also proposed to calibrate the items and obtain the participants' latent scores.

Method

Participants

A sample of 724 psychology students (76% women and 24% men; $n_1 = 132$, $n_2 = 121$, $n_3 = 112$, $n_4 = 117$, $n_5 = 127$, $n_6 = 115$ for each subtest) was selected, aged from 18 to 30 years old (M = 19.51, SD = 1.69). Of them, 169 students agreed to be assessed also in general mental ability, 271 in working memory, and 145 agreed to perform some tasks of attention control. The sampling was intentional and was carried out in the *Universidad Autónoma de Madrid*. Participation was voluntary.

Instruments

Item pool development. Three experts in psychological evaluation and psychometrics developed a total of 54 progressive matrix items, following the taxonomy rule proposed by Carpenter et al. (1990). According to Primi (2001), item difficulty was manipulated by altering the number of attributes present in each item and varying the number of rules involved in their resolution (i.e. constant per row, quantitative pairwise progression, figure addition or subtraction and distribution of two or three values). As shown in Figure 1, each item is a 3 × 3 matrix with the lower square empty. The response format is constructed, so that once respondents have figured out which rules are followed by an item, they must draw their own response in the lower square, selecting different colors and patterns for the 16 empty cells to compose their answer. In this case, respondents should guess how changes the background of each row and how it affects to the figures in each square. Then they have to notice the parallelism between these rules and the new situation (i.e. the empty square) to, finally, apply correctly the rules to compose a new solution that fit in the matrix, drawing a background of vertical lines and a 2-cell figure with vertical and horizontal lines.

Items are then scored from 1 to 5, depending on the number of correctly filled cells: 1 (7 or fewer right cells), 2 (from 8 to 11 right cells), 3 (12 or 13 right cells), 4 (14 or 15 right cells) and 5 (the correct answer, 16 right cells).

The items were designed by triads, so for each original item designed, two more clone-items were created by modifying some accessory features (i.e., color, shape) of the original item, but keeping the same rules needed to find the right response. Therefore, the rules needed to solve an original item of a triad are the same rules needed to solve their clones. Eighteen original items were developed and, based on them, 36 new items (the clones) were also developed, with a total number of 54 items in the pool. These items were originally developed to construct a test based on automatic item generation.

Subtest development. In order to avoid the inclusion of two or more items following the same rules in the same test, a counterbalanced anchoring design (von Davier, Holland, & Thayer, 2004) was carried out for the application and calibration of the item pool. Six subtests were then assembled, each one comprised of 18 items from different triads.

Additional measures. With the aim of gathering some validity evidence, several measures related to fluid intelligence were also assessed. General mental ability was measured by the Advanced Progressive Matrices test (APM; $\alpha = .71$), the subscales R (a logical series test; $\alpha = .74$) and V (a vocabulary test; $\alpha = .76$) of the Spanish version of the Primary Mental Ability test (PMA; Cordero-Pando, 1984), and the subscales of abstract reasoning (AR; $\alpha = .66$), numeric reasoning (NR; $\alpha = .63$), and verbal reasoning (VR; $\alpha = .61$) from the Spanish adaptation of the Differential Aptitudes Test (DAT-5; Corral & Cordero-Pando, 2006). Working memory was evaluated with the Spanish adaptation of the reading span (Elosúa, Gutiérrez, García-Madruga, Luque, & Gárate, 1996),

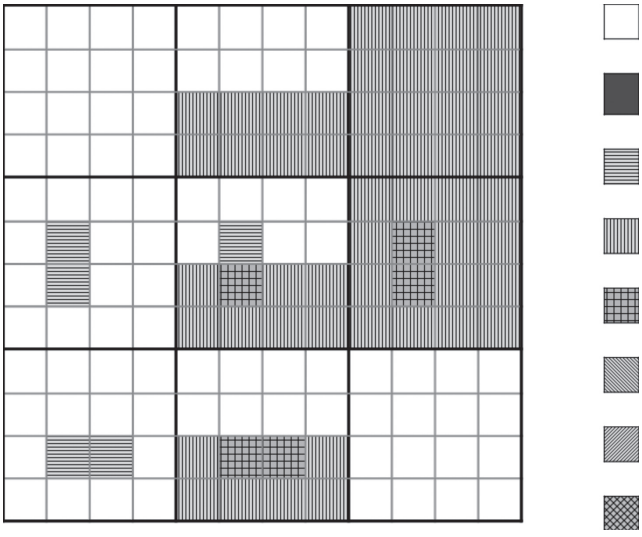

Figure 1. *Example item*

the computation span (Ackerman, Beier, & Boyle, 2002), and the dot matrix task (Miyake, Friedman, Rattinger, Shah, & Hegarty, 2001).

Procedure

The sampling was carried out in three sessions. Participants were randomly assigned to one of the fluid intelligence subtests in the first session. Then, they performed the tasks of working memory and the rest of the tests in a second and third sessions, respectively. They had two hours per session to perform the experimental tasks and fulfill the questionnaires.

Data analysis

The analyses were conducted using the free statistical software *R*. Item descriptive statistics and the internal consistency of the subtests were obtained with the *psych* package (Revelle, 2015). Validity evidence, based on the relations with other variables, was assessed by correlating the additional measures with the raw scores in each subtest. The *mirt* package (Chalmers, 2012) was utilized to calibrate the item pool under the graded response model (Samejima, 1969). This package was also employed to study the dimensionality of each subtest separately with the M_2 statistic, a limited-information goodness-of-fit statistic developed to assess the fit of overall IRT models utilizing marginal residuals of item pairs or triplets (Maydeu-Olivares & Joe, 2006). This statistics is interpreted as a χ^2 test; if the M_2 results no significance, it means good overall fit.

Results

Descriptive analysis and reliability. The item pool presented different difficulty levels, with 24 relatively easy items –with a mean of 4 or higher–, 22 items of moderate difficulty –with a mean between 3 and 4–, and 8 items of higher difficulty –with means between 2 and 3. Thus, the standard deviation of the items (around 1) was indicative of an adequate heterogeneity in the participants' responses. Regarding internal consistency, the Cronbach indices of the subtests were $\alpha_1 = .82$, $\alpha_2 = .79$, $\alpha_3 = .70$, $\alpha_4 = .72$, $\alpha_5 = .77$, and $\alpha_6 = .82$.

Dimensionality and IRT modeling. The subtests were calibrated separately under the graded response model, and the M_2 statistic was computed for each subtest. This index showed an adequate fit in each subtest when the items were calibrated under the graded response model ($M_{2\ Subtest1}(88) = 80.16$, $p = .712$; $M_{2\ Subtest2}(85) = 75.64$, $p = .756$; $M_{2\ Subtest3}(91) = 93.32$, $p = .413$; $M_{2\ Subtest4}(89) = 81.81$, $p = .693$; $M_{2\ Subtest5}(87) = 60.14$, $p = .987$; and $M_{2\ Subtest6}(87) = 81.24$, $p = .654$). Therefore, each subtest was evaluating a single dimension.

The restrictions of item administration were also studied. Although it is expected that items comprising the same triad will have similar statistical properties, Glas and van der Linden (2003) pointed out that there is some variability between originals and clone-items which should be considered. For this purpose, we tested whether the item-parameter estimates are equal for items comprising the same triad (for further details, see Martín-Fernández, Ponsoda, Olea, Shih, & Revuelta, 2015). A total of 14 original-items and 10 clone-items of their triads were released, increasing the item pool and improving the test information.

Several nested models were then proposed to calibrate the entire item pool by applying appropriate linking procedures (von Davier et al., 2004), considering the relations among the items from a same triad. Four nested graded response models were proposed: (1) an unconstrained model, (2) a model fixing the slope (a_j) and the threshold (b_{jk}) parameters of the not-released clones of the same triad to the same value, (3) a model with only the threshold (b_{jk}) parameters of the not-released clones fixed to the same value, and (4) a model fixing only the slope (a_j) parameters of the not-released clones to be equal.

Compared with the unconstrained model, only the slope-fixed one (model 4) showed that there were no significant differences between the two models in terms of fit. So finally the more parsimonious model was kept, forcing the *a*-parameters of some clones of the same triad to be equal (see Table 1).

Relations with other variables. The raw scores in the subtests were positively related with the other general mental ability measures (see Table 2). They showed a strong relation with the Raven test and the other general mental ability measures, except for the V scale of the PMA –the vocabulary test. The subtest scores also showed a positive and strong relation with the working memory tasks.

Discussion

This first study showed that the psychometric properties of the progressive matrix item pool are adequate to evaluate fluid intelligence. On the one hand, the internal consistency and the dimensionality of the six subtests indicated a good reliability of the measure, and that each subtest is measuring a single latent

Table 1
Model comparisons

Model	Parameters	log-likelihood	G^2	df	p	AIC	BIC
1	258	-13571.55				27659.1	28841.98
2	205	-13643.13	143.156	42	<.001	27718.26	28708.57
3	214	-13638.71	134.312	33	<.001	27727.41	28758.99
4	248	-13574.80	6.509	9	0.688	27647.61	28789.22

Note: Model 1: unconstrained model, Model 2: slopes and thresholds fixed, Model 3: thresholds fixed
Model 4: slopes fixed

Table 2
Correlations among the subtest scores and the additional measures

	FIMT
APM	.47**
DAT5 - AR	.52**
DAT 5 - NR	.23**
DAT 5 - VR	.39**
PMA – R	.28**
PMA – V	.12
Reading span	.29**
Computation span	.30**
Dot matrix	.42**

**p<.001

construct. On the other hand, the close relation between the score of the participants in the subtests and the general mental ability measures, in particular with the Raven test, are highlighted. As expected, the verbal reasoning measures were the less related to the subtest scores (Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004). The correlations with the working memory measures were also congruent with previous studies (Colom et al., 2007).

In conclusion, this study provides a fully operational item pool of fluid intelligence, ready to allow a multistage test to be assembled with it.

Study 2

The second study explores different multistage structures by automated test assembly (ATA) in order to obtain the most informative form of the Fluid Intelligence Multistage Test (FIMT) for the higher latent trait values.

Method

Participants

The same sample of 724 psychology students was utilized to find the optimal multistage test structure.

Instruments

Item pool. The operational item pool of fluid intelligence of the Study 1, which eliminated some restrictions in the administration of the original and clone-items, was employed in this study.

Procedure

Two multistage structures were tested in this study (see Figure 2). Both structures were divided into three stages, with one general routing test in the first stage followed by two more stages with two (structure 1-2-2) or three modules (structure 1-3-3) per stage. Hence, respondents were first evaluated with a module of general difficulty –the routing test– and then they were sent on to a new module of variable difficulty, depending on the performance of each respondent on the routing test. Once the second stage had finalized, respondents were again sent on to a new module of a difficulty according to their performance in the previous stages. Finally, when the third stage was completed, the latent trait θ was estimated for each respondent.

When setting the ATA constraints, two strategies were followed in order to increase the information of the test for examinees with a high ability level. First, items could be repeated across the modules of different stages as long as the same item was not presented twice in the same path of the multistage test. And second, the high difficulty module of the third stage was five items longer than the rest of the modules. Therefore, the ATA constraints were formulated as follows:

Routing test

$$\max y \tag{1}$$

subject to:

$$\sum_{j=1}^{J}\sum_{k=1}^{K} I_j(\theta_k)x_j \geq y, \theta_k \in [-1.75, 1.75] \tag{2}$$

$$\sum_{j\in V_c} x_j \leq 1 \tag{3}$$

$$\sum_{j=1}^{J} x_j = N_1 \tag{4}$$

$$x_j \in \{0,1\} \tag{5}$$

$$y \geq 0 \tag{6}$$

The ATA first needs to convert the construction task into a linear optimization problem. The objective of the problem is expressed in (1): maximize the constant *y*. Constraint (2) means that the sum of the information of the selected items in the interval between θ = -1.75 and θ = 1.75 must be equal or higher than *y*. Thus, if *y* must be maximized, then the test information is the maximum possible. As expressed in (5) x_j is a dichotomous variable indicating whether or not item *j* is included in the test. Thereafter, the combination of items that provides more information to this theta interval is selected.

Besides the information constraint, (3) express that more than one clone-item of a certain triad cannot be presented

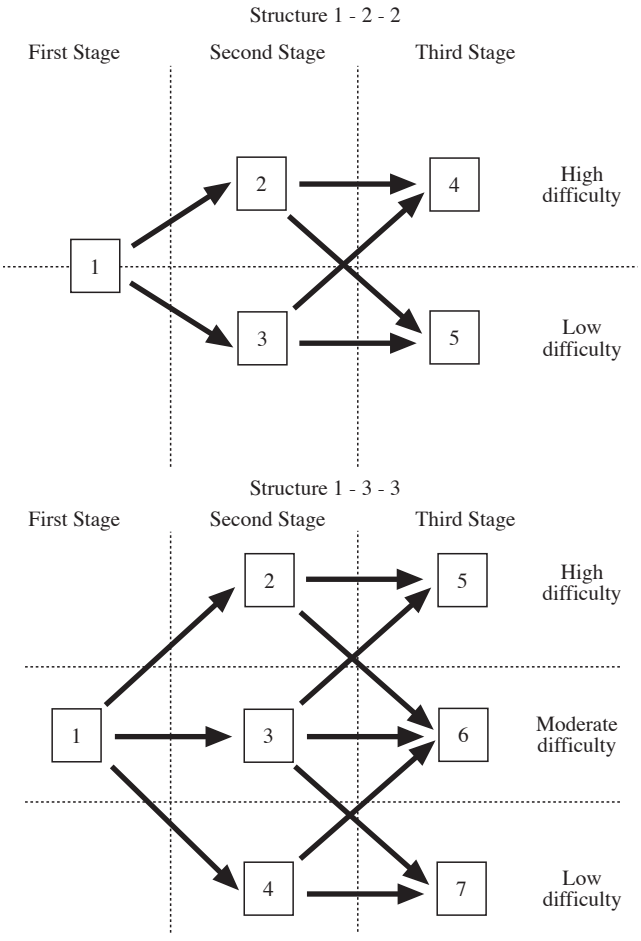

Figure 2. *FIMT tested multistage structures*

simultaneously (V_c denotes the group of j items of the pool that cannot be presented in the same test), and (4) fix the length of the module to N_1. At least, constraint (6) is imposed, preventing y from taking negative values.

The second and the third stages introduce minor changes to the previous constraints:

Second Stage

$$\sum_{j=1}^{J}\sum_{k=1}^{K} I_j(\theta_k)x_j \geq y, \theta_k \in [\theta_1, \theta_h] \quad (7)$$

$$\sum_{j=1}^{J} x_j = N_2 \quad (8)$$

$$\sum_{j \in S_{prev}} x_j = 0 \quad (9)$$

Constraints of the second stage are very similar to the routing test ones with a notable exception: the θ interval to maximize changes from module to module. The objective function remains the same, maximize y, in order to obtain the maximum information possible for a certain interval of θ (7). In the 1-2-2 structure, the intervals of the modules are: $\theta_l = -1.75$ and $\theta_h = 0$; and $\theta_l = 0$ and $\theta_h = 1.75$. In the 1-3-3 structure, however, the intervals to maximize information are: $\theta_l = -1.75$ and $\theta_h = -1$; $\theta_l = -1$ and $\theta_h = 1$; and $\theta_l = 1$ and $\theta_h = 1.75$. The constraints (8) and (9) fix the length of each module to N_2, and impede the inclusion of items presented in the *Sprev* set, which are the items already included in the path (in this case, the items of the routing test).

Third Stage

$$\sum_{j=1}^{J} x_j = N_3, N_3 = c \text{ if } \theta_h \neq 1.75$$
$$N_3 = c + 5 \text{ if } \theta_h \neq 1.75 \quad (10)$$

In the third stage, the constraints again remain almost equal with respect to the second stage. The only relevant change is the one expressed in (10), which denotes that the length of the maximum difficulty module must be 5 items larger than the length of the rest of the modules of the stage.

Data analysis

To determine which multistage structure was most appropriate, the number of actual paths of each structure and the test information provided by the paths of each structure was considered. To summarize the information of each path, the sum of $K = 121$ discrete points across the interval $\theta = -3$ and $\theta = 3$ for all the J items comprising each path were taken:

$$I(\theta)_{path} = \sum_{j=1}^{J}\sum_{k=1}^{K} I_j(\theta_k), \textit{for } \theta = (-3, -2.95, -2.90, \ldots, 3)$$

Analyses were conducted with the *lp_SolveAPI* package (Konis, 2014) for the ATA, and the *mirt* package (Chalmers, 2012) to compute the test information function.

Results

Test structure. The average information across the $\theta = -3$ and $\theta = 3$ interval for the 1-2-2 configurations was always higher than the same configurations for the 1-3-3 structure (see Table 3). Regarding the number of different paths, the 1-2-2 structure always presented the four possible paths of its structure. Nonetheless, in 1-3-3, we found one or two repeated paths among the seven possible itineraries of this multistage structure.

The differences between the two structures were minimal but, given that these little discrepancies favored the 1-2-2 structure, the simplest structure was kept. Between the different lengths considered for the modules of the stages, the one comprised of five items for the first stage routing test, six items for the modules of the second stage, and four items for the modules of the third stage was chosen as the final FIMT because it was the most informative for the latent trait interval.

Information function. As shown in Figure 3, each curve corresponds to a different path in the multistage structure. Besides the four paths of the multistage test, two more curves were added to the figure (i.e. large discontinued lines with two points in between), corresponding to the information provided by a 15- or 20-item test comprised of randomly selected items. In fact, the information of the different FIMT paths was always more informative than the random test, even with 5 items less.

The low/low path is comprised of the items of the routing test, the low difficulty module of the second stage, and the low difficulty module of the third stage. That is why this path resulted especially informative for the lower levels of the latent trait. Therefore, the high/low path was the one followed by the respondents who did the routing test well but failed in the high difficulty module of the second stage and were sent on to a lower difficulty module of the third stage. This path provided less information for the low levels of θ than the previous path but the drop of the curve was not so sudden in intermediate and high levels of the latent trait. The low/high and the high/high paths were administered to respondents with a good performance in the second stage, and are comprised of 20 items instead of 15. That is why the information values were in general higher than in the other paths. The low/high path was more informative for low and moderate θ values than the high/high path. However, the high/high path resulted slightly more informative for high levels of the latent trait.

Table 3
Average information of each multistage configuration

Module length			Structure 1-2-2	Structure 1-3-3
S1	S2	S3*	Mean I(θ)	Mean I(θ)
7	5	3	614.75	587.26
7	4	4	617.05	587.10
7	3	5	617.05	587.73
5	6	4	622.21	589.19
5	5	5	618.85	589.66
5	4	6	617.74	589.13
3	7	5	619.67	590.05
3	6	6	619.67	590.35
3	5	7	619.67	589.52

Note: S1 = Stage 1; S2 = Stage 2; S3 = Stage 3.
*: Stage 3 is 5 items longer for the high difficulty modules

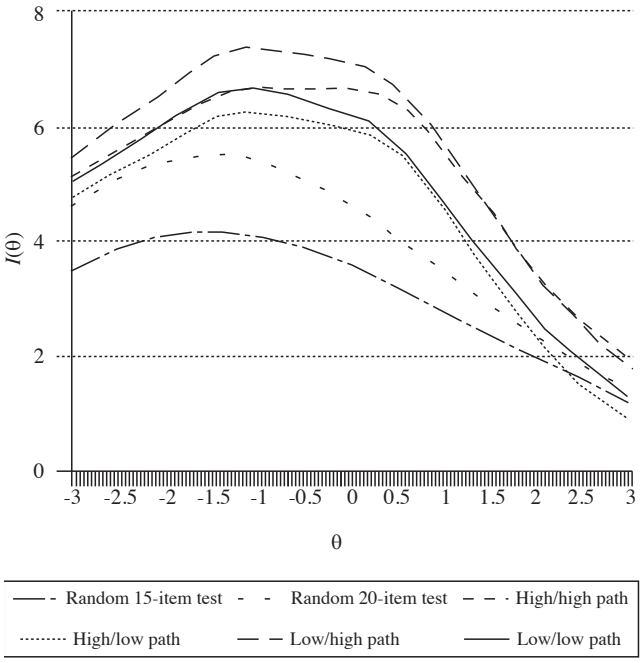

Figure 3. *FIMT information function of the paths and two random tests*

Discussion

The 1-2-2 structure was the best adapted to the characteristics of the item pool (i.e., limited number of items, administration restrictions). It demanded a lower number of items, and the items comprising the modules were the most informative for the different latent trait levels considered. However, the 1-3-3 structure needed more items to be assembled, which consumed the most informative items in the early stages, and due to the administration constraints, fewer informative items were available for the last modules. The other drawback of the 1-3-3 structure was that it did not use all seven possible paths because some of the items are repeated, comprising two paths with the same items but presented in a different order.

Thanks to ATA, an optimal multistage structure was achieved for the FIMT, granting the test precision, and adapting its difficulty to the examinees' ability.

General Discussion

The purpose of this research was to develop a multistage test to assess general mental ability, focusing on examinees with high ability levels.

Through the current studies, the psychometric properties of the initial item pool have been analyzed and different multistage structures have been tested to finally reach a fully operational adaptive multistage test. This is one of the key aspects of this research, because it constitutes an empirical application with a small item pool of a procedure normally employed with large item pools as a heuristic to decide which items should comprise the final form of a given test (Swanson & Stocking, 1993).

The multistage format of the FIMT allows some interesting options for test-based assessments. On the one hand, multistage testing assembled via ATA allowed an optimal treatment of the items comprising the final test, maximizing the information of the measure in comparison with an arbitrary test assembly criterion. On the other hand, the technical implementation for multistage tests is far simpler than for CAT, especially when the tests are relatively short and the estimates of each response pattern of the multistage test can be previously determined, making the assessment more computationally efficient, as it is not necessary to run any estimation method during test administration.

Another interesting feature of the FIMT is the constructed response format of the items, which denies respondents the possibility of guessing the right answer from a given set of alternatives and also impedes obtaining a high score in the test with low elaborated responses.

This research is not without limitations. Although the dimensionality of the subtests were assessed, it is necessary to study the dimensionality of the paths of the final multistage test. Moreover, predictive validity evidence should also be gathered in further studies, relating the scores in the different paths to other variables, like job performance.

The response format of the items also invites one to try other treatments of the answers. Although a polytomous response was coded and the graded response model was picked, other options are available. Other cut-offs could be established to demarcate the response categories of the items. Moreover, considering the total number of cells composing the empty square of each matrix, the continuous response model (Samejima, 1974) could be an interesting alternative to calibrate the items and score the examinees. The logistic family models could also be considered if the responses are coded dichotomously. Furthermore, other theoretical models based on the cognitive processes (Kunda, McGreggor, & Goel, 2013) could be considered to score the items.

Although the items were developed to assess high ability levels, the information provided by the FIMT for very high ability values (above 2) falls abruptly. This is due most likely to the low discrimination parameters of the most difficult items in the item pool. One way to improve this feature of the test is to develop new items requiring either more rules to be correctly resolved or more attributes present in the items (Primi, 2001).

Another question worth considering is whether the FIMT is really a multistage test. Usually, once an item has been assigned to one of the multistage modules, it cannot be selected again. Besides, the length of the modules of the same stage tends to be the equal. If these constraints to the ATA had not been applied, then the information function of the resulting paths would have dropped substantively, leading to quite imprecise person-parameter estimates. However, the FIMT has a test structure divided into stages with modules of different difficulties, allowing some adaptability between the test and the examinee's ability. It is precisely this adaptive nature and the control assumed over the contents present in each module of the test what makes the FIMT the best multistage test that can be assembled with this item pool.

Acknowledgements

Authors would like to thank to the *Instituto de Ingeniería del Conocimiento* and to the *Cátedra "Modelos y Aplicaciones Psicométricos"* their support for the present research.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General, 131*, 567-589.

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131*, 30-60.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29.

Colom, R., Abad, F. J., Quiroga, M. A., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence*, *36*(6), 584-606.

Colom, R., Escorial, S., Shih, P. C., & Privado, J. (2007). Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents. *Personality and Individual Differences*, *42*(8), 1503-1514.

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence*, *32*(3), 277-296.

Cordero Pando, A. (1984). *PMA: Aptitudes Mentales Primarias 6ª ed. rev.* [PMA: Primary Mental Aptitudes 6th edition]. Madrid: TEA Ediciones.

Corral, S., & Cordero Pando, A. (2006). *DAT-5: Test de Aptitudes Diferenciales Versión 5* [DAT-5: Differential aptitude test]. Madrid: TEA Ediciones.

Diao, Q., & van der Linden, W., J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement*, *35*, 398-409.

Elosúa, M. R., Gutiérrez, F., García-Madruga, J. A., Luque, J. L., & Gárate, M. (1996). Adaptación española del "Reading Span Test" de Daneman y Carpenter [Spanish standardization of the Reading Span Test]. *Psicothema, 8*, 383-395.

Glas, C. A., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*(4), 247-261.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*(2), 44-52.

Konis, K. (2014). *lpSolveAPI: R Interface for lp_solve version 5.5.2.0. R package version 5.5.2.0-14*. Retrieved from https://cran.r-project.org/web/packages/lpSolveAPI/index.html.

Kunda, M., McGreggor, K., & Goel, A. K. (2013). A computational model for solving problems from the Raven's Progressive Matrices intelligence test using iconic visual representations. *Cognitive Systems Research*, *22*, 47-66.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*, 229-249.

Martín-Fernández, M., Ponsoda, V., Olea, J., Shih, P.C. & Revuelta, J. (2015). *Test multietapa de inteligencia fluida* [Fluid Intelligence Multistage Test]. Retrieved from: http://www.iic.uam.es/catedras / map/publicaciones.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713-732.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General, 130*, 621-640.

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence*, *30*(1), 41-70.

Revelle, W. (2015). *Procedures for personality and psychological research*. Evanston, Il: Northwestern University.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. New York, NY: Psychometric Society.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, *39*(1), 111-121.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. *Advances in the Psychology of Human Intelligence*, *2*, 47-103.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*(2), 151-166.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer.

Wang, X., Fluegge, L., & Luecht, R. (2012). *A large-scale comparative study of the accuracy and efficiency of ca-MST*. Paper presented at the annual meeting of the National Council on Measurement en Education, Vancouver, BC, Canada.

Yan, D., von Davier, A. A., & Lewis, C. (2014). *Computerized multistage testing: Theory and applications*. New York, NY: CRC Press.