JOURNAL OF APPLIED MEASUREMENT, *18*(2),

Development of an Item Bank for the Assessment of Knowledge on Biology in Argentine University Students

Marcos Cupani

Tatiana Castro Zamparella

Gisella Piumatti

Grupo Vinculado

Centro de Investigaciones y Estudios sobre Cultura y Sociedad (CIECS), CONICET. Facultad de Psicología, Universidad Nacional de Córdoba.

The calibration of Item Banks provides the basis for computerized adaptive testing that ensures high diagnostic precision and minimizes participants' test burden. This study aims to develop a bank of items to measure the level of Knowledge on Biology using the Rasch Model. The sample consisted of 1219 participants that studied in different faculties of the National University of Córdoba (mean age = 21.85 years, *SD* = 4.66; 66.9% are women). The items were organized in different forms and into separate subtests, with some common items across subtests. The students were told they had to answer 60 questions of knowledge on biology. Evaluation of Rasch model fit ($Zstd \leq \pm 2.0$), differential item functioning, dimensionality, local independence, item and person separation (>2.0), and reliability (>.80) resulted in a bank of 180 items with good psychometric properties. The bank provides items with a wide range of content coverage and may serve as a sound basis for computerized adaptive testing applications. The contribution of this work is significant in the field of educational assessment in Argentina.

Requests for reprints should be sent to Marcos Cupani, Laboratorio de Psicología de la Personalidad, Facultad de Psicología, Universidad Nacional de Córdoba, Ciudad Universitaria, Córdoba 5000, Argentina. e-mail: marcoscup@gmail.com.

Science is shaping the lives of people in a fundamental way. Individuals, groups, and nations are increasingly seeking to strengthen scientific capacity, in the hope of promoting social and personal well-being (Feder, Shouse, Lewenstein, and Bell, 2009). Somehow, Biology is the most rigorous of all sciences; on the one hand, because living systems are too complex and, on the other, because Biology is an interdisciplinary science that requires the knowledge of Chemistry, Physics, and Mathematics (Campbell and Reece, 2007). For these reasons, efforts to improve scientific capacity are often focused on educational institutions and on improving strategies, such as science courses, teacher training, and educational measurement systems.

Educational measurement is generally regarded as empirical, quantitative and, on a large scale, as having the main purpose of controlling educational systems (Long, Wendt, and Dunne, 2011). These evaluation systems can give the institutions the opportunity to measure their own progress year after year in compliance with the standards set by government agencies (Garbanzo Vargas, 2007) and can also assess the quality of educators' training (Popham, 2001). In addition, these evaluation systems are used for university admission, certification, monitoring, and diagnosis of student's learning (Eggen, 2011). In other words, the evaluation of academic performance is a very important issue for parents, teachers, and the Government (Novak, Mintzes, and Wandersee, 2000). An accurate measurement plays a very important role in the evaluation of academic success and, therefore, the development of analytical methods has advanced greatly in recent years (Törmäkangas, 2011).

In everyday university life, academic performance is assessed through teachers' perception and judgment. The magnitude of this tentative measurement is typically only ordinal and its results are open to speculation. The use of standardized instruments can minimize students' misclassification and can enable the measurement of academic performance with an interval scale. Although most researchers in the field of education use the classical test theory (CTT), at present, the item response theory (IRT) and the Rasch model (Rasch, 1960) have gained popularity (Hambleton, 2000). This is because such models can provide invariant measures, regardless of the instruments used and the individuals evaluated (Hambleton, 1985).

The Rasch model provides a comprehensive and detailed methodology that can evaluate the psychometric properties of an instrument at item level (Messick, 1994). Rasch analysis provides more information about the ability of a person because it focuses on the difficulty of the items, rather than the number of items that each participant answers correctly. From a Rasch perspective, the ability of a person interacts with the item difficulty to obtain a score for each subject in the measurement (Linacre, 2002). To analyze the items, the Rasch model first converts the ordinal data from an instrument into interval data, thus fulfilling one of the prerequisites of any measurement (Wright and Linacre, 1989). Then, this psychometric model enables the evaluation of several characteristics, such as the model fit level, the item difficulty and hierarchy, the reliability of persons and item, and the differential item functioning (DIF). In short, the Rasch model can help test developers and science educators to improve the validity, reliability, and efficiency of educational instruments (Bond, 2003).

At present, the use of computers combined with IRT and Rasch model enables the building of Computerized Adaptive Testing (CAT). With CAT, the items given are adapted to the level of competence that the subject is showing according to their responses to previous items (Barrada, Olea, Ponsoda, and Abad, 2010). A prerequisite is to have a calibrated item bank (IB); that is, a set of items that make it possible to measure the latent variable unidimensionally by considering the different dimensions that integrate this variable (Wright and Bell, 1984). The capacity of CAT to provide a fast and reliable estimate of a person's skill level is based on the quality of the items in the system (Lai, Dineen, Reeve, Von Roenn, Shervin, McGuire and Cella, 2005). Moreover, a calibrated IB can be a good starting

point for teachers to select a specific set of items for particular assessments as well as parallel tests.

Therefore, this study aims to develop an item bank to measure knowledge on Biology. Through this calibrated item bank an assessment and a classification of students will be performed. In the case of the assessment, this test will give an estimate of the level of knowledge acquired in the domain of Biology. In the case of the classification, this IB will enable educational decision making, for example, admission to a major or an educational grant. On the other hand, a calibrated IB using the Rasch model is the basis for the development of the CAT to measure knowledge on Biology.

Method

Participants

The sample consisted of 1,219 students, 815 female persons (66.9%) and 404 male persons (33.1%), between 17 and 58 years old (M = 21.85, SD = 4.66) that studied in different faculties of the National University of Córdoba (UNC): Medicine (25.1%), Biology (15.7%), Nutrition (15%), Nursing (13%), Speech Therapy (11.2%), Medical Technology (10.2%), Psychology (8.5%), and Agricultural Engineering (1.3%). As regards their academic year, 395 participants completed the Introductory Level (32.4%); 443, the first academic year (36.3%); 293, the second academic year (24%); 69, the third academic year (5.7%); and 19 the fourth academic year (1.6%). The participants answered the following forms: Form A Level I (n = 304), Form B Level I (n = 311), Form A Level II (n = 301), and Form B Level II (n = 303).

Instrument

Item Bank on Biology, Levels I and II. For a test to measure a specific domain accurately and reliably and to provide validity evidence to support the inferences made with the test scores, systematization, and organization of the activities to develop are required (Downing and Haladyna, 2006). The activities carried out are described below.

Content analysis and specification table. In order to define the content to be measured, 55 syllabuses belonging to seven majors (Biology, Medicine, Nutrition, Medical Technology, Nursing, Speech Therapy, and Psychology) related to the Natural Sciences (NS) and the Health Sciences (HS) at UNC were collected. Each syllabus was organized in a spreadsheet (Excel) by syllabus, syllabus academic year (n = 4), didactic units (n = 519) into which each syllabus is divided (general contents) and topics (specific contents). From the general and specific contents, six specialists established *conceptual units*. These were processed by the authors of this study using a frequency analysis to show in descending order what the most important ones (occurring more frequently) are.

This information was organized on three levels: Level I, contents taught in first year; Level II, contents taught in second year; Level III, contents taught in third year; and Level IV, contents taught in fourth year. Then a group of five experts evaluated the representativeness of the selected information. Once the most representative contents were selected, a specification table with 100 written questions per level was created. These questions were distributed evenly in different contents and cognitive categories (knowledge, understanding and application).

Item writing and Development. The writing of the items was done by ten professionals in NS and HS. They received special training on guidelines for the creation of multiple-choice items (Haladyna, Downing, and Rodriguez, 2002). Each professional received the specification table where the number of questions by content was specified. In all, 532 initial questions for the four Levels were written. These questions were organized into cards with a unique identification code, a related concept, the type of cognitive category evaluated, the right choice and a justification of why each choice is either right or wrong, the bibliographic source used for writing the question, and who the author of the question was. A space to categorize the level of difficulty was also created. Subsequently, these cards were given to five judges who evaluated the selected

specific content for each item. In addition, the judges rated the questions according to their level of difficulty as easy, medium, and high. Some questions were discarded and the final pool was comprised of 487 items.

Test design, assembly, and production. The items were organized in different forms in order to be calibrated. For Level I, Forms A (60 items) and B (60 items) were prepared; for Level II, Forms A (60 items) and B (59 items) were prepared; for Level III, Forms A (60 items) and B (60 items) were prepared; and for Level IV, Forms A (60 items) and B (60 items) were prepared. The item distribution in each form was performed by increasing difficulty level and different contents. Moreover, in each form anchor items (20 to 40) and free items (20 to 40) were established. The options of the correct answers varied randomly in location. On the other hand, the same format for all forms was established: (a) a question paper in double format to facilitate the reading and (b) an answer protocol to organize the scores of the evaluated participants with set spaces for choosing their answer (A, B, or C). We consider that three options are enough because the effort for developing a fourth option (the third plausible distractor) is probably not worth it. It is very unlikely that item writers can write three distractors that have item response patterns consistent with the idea of plausibility (Haladyna and Downing, 1989).

Procedure

The administration of the test was done collectively, in a regular class schedule and under the supervision of teachers assigned to the class schedule. Prior to administration, the students were told they had to answer 60 questions of knowledge on biology and that all questions only had one correct option. They were also recommended to try to answer all the questions and, should they consider the question was oblivious to them, not to answer. After this statement, the students were given an informed consent and the material to be read and answered. Answering the entire test took between 40 and 60 minutes.

Data Analysis

All analyses were performed with the Rasch model, which ensures that all parameters of people and items are specific locations in a single latent variable and can be expressed in the same unit scale (*logit*), thus making it possible to establish objective comparisons. The calibration plan consisted of the following steps.

Step A. Unidimensionality and local independence. Unidimensionality was assessed by the NOHARM (Normal Ogive Harmonic Analysis Robust Method) program version 4.0, which allows to evaluate the relationship between the nonlinear factor analysis and the normal ogive model according to the dimensional and multidimensional adjustment of the normal ogive model (De Ayala, 2013). NOHARM produces a residual matrix to evaluate the fit of the model. The software provides the root mean square residual (RMSR) where values close to zero represent an adequate fit to the model. If the RMSR is greater than the residual standard error ($4 / \sqrt{N}$), it indicates that the model does not fit well (Fraser and McDonald, 2002). A second index of adjustment is the Tanaka's Index (1993) for goodness of fit (GFI). Mc Donald (1999) suggests that a score of 0.90 is an acceptable value and that an index of 0.95 indicates a good fit. The assumption of local independence was evaluated by inspecting the residual matrix and the covariance matrix, where values less than 0.025 and 0.25, respectively, are expected.

Step B. Rasch model fit. Three analyses were performed: global adjustment of data, items adjustment, and people adjustment. The first checks whether the data matrix, in a broad sense, is in line with what is predicted by the model. The items fit enables to study each of these items independently. In addition, with the adjustment of the persons, the persons who have responded inconsistently to the theoretical formulation can be identified. In order to check if there is a fit between the data and the model, we follow different procedures, the most frequently used in the RSM being the residuals analysis (Infit and Outfit). WINSTEP reports two Infit statistics (*Mnsq* and *Zstd*) and

two Outfit statistics (*Mnsq* and *Zstd*). Following the criteria proposed by Linacre (2009), the values of the Mnsq statistics in the interval between 0.6 and 1.3 would have an acceptable fit. An acceptable Zstd statistics fit would fluctuate between values equal to or higher than +2 and equal to or lower than −2 (Bond and Fox, 2007). The point-biserial coefficients (*rpbis*), which are useful for diagnosing errors when coding the items were also calculated (values that are negative or equal to zero indicate items with response patterns that contradict the variable).

Step C. Separation and reliability. The items should be separated in difficulty levels well enough to identify the meaning and significance of the latent variable (Wright and Stone, 2003). The person separation index indicates how effectively the instrument can discriminate against persons on the measured variable. A useful set of items must define at least three layers of persons (for example, high, moderate, and low levels of knowledge). A separation index greater than 2 is considered adequate (Bond and Fox, 2007) as well as a reliability associated with a person separation index of 0.80 (Gauggel et al., 2004).

Steps D. Differential item functioning (DIF). DIF analyses were performed according to participants' gender and age. An item has DIF when the probability of a correct answer depends not only on the level of the person on the trait intentionally measured by the test. For the DIF analysis by age, the participants were ranked as "young" and "adult" by calculating the median. To apply the DIF, pairwise analyses where the significance level was set at $\alpha < 0.01$ were performed and it was considered that the DIF contrast must be ≥ 0.5 *logits* (Linacre, 2009). In addition, we estimated the size of the effect based on the sample (DIF contrast / *SD* measure). We used *t*-test with Welch-approximation of the degrees of freedom on Rasch person estimates using Winsteps. The *t*-test is a two-sided test for the difference between two means (i.e., the estimates) based on the standard error of the means (i.e., the standard error of the estimates).

Step E. Specific objectivity. An analysis of the specific objectivity of the anchor items was performed. This is one of the most important properties of the Rasch model and it refers to the fact that a measure can only be considered valid and generalizable if it does not depend on the specific conditions with which it has been obtained. One of the main procedures recommended to analyze the data fit model is to contrast this property empirically (Hambleton, Swaminathan, and Rogers, 1991). In this study, to analyze the invariance of the item parameters, (i) the items of Forms A and B were unified on the same basis, (ii) this basis was divided randomly into two, (iii) the parameters of anchor item difficulty were estimated, and (iv) a simple linear regression between the parameters obtained was carried out. The expected values of the correlation between the two sets of parameters, the intercept and the slope of the line indicating a perfect fit value would be 1, 0, and 1, respectively (Prieto and Delgado, 2003).

Results

Level I Items

Form A

Step A. The RMSR value (0.0146) is lower than the typical residual estimated error (0.23), which indicates that the assumption of unidimensionality is true. Tanaka's goodness of fit index, however, was 0.84 and did not exceed the proposed cut-off point. Furthermore, the values of the covariance matrix did not exceed the cut-off value of 0.25. Moreover, it was seen that 7% of the residues of all items were lower than 0.025. This could indicate that there may be one or more factors that explain the remaining variance (Yen, 1993). In this sense, as this is a test that measures a general factor composed of more specific factors (knowledge in biology was defined in terms of knowledge about *cell*, *states of the matter*, *macromolecules*, *living organisms*, *the origin of life*, *branches of biology*, and *chemical reactions*, among others), it is expected to obtain a complex factorial structure (Tate, 2003), and that performance on an item related to the performance of another item that requires knowledge, skills, and similar capacities (Yen, 1993).

Step B. All 60 items showed an adequate fit to the model (*Mnsq* ≤ 1.3 or ≥ 0.6). Ten items, however, showed inadequate fit to the model (*Zstd* ≤ ±2.0) and *rpbis* values that were negative or close to zero. These items were eliminated and the model was recalculated with 50 items. The fit of these was satisfactory for both *Mnsq* and *Zstd*. The measure of difficulty (δ_i) of the items varied between $-3.40 \leq \delta_i \leq 1.91$, with a mean of 0.00 (*SD* = 0.84). The analysis of the persons' adjustment reflects that 95% of the response patterns adjusted to the model (*Mnsq* ≤ 1.3 or ≥ 0.6). The skill levels varied between $-2.45 \leq \theta \leq 1.70$, with a mean of −0.45 (*SD* = 0.68).

Step C. The item separation (6.17) and item reliability (.97) were satisfactory, which indicates that the sample used is large enough to confirm the hierarchy of item difficulty (construct validity) of the instrument (Linacre, 2009). On the other hand, the values of person separation (1.86) and person reliability (.78) were considered acceptable, although the need to cover some skill levels with other questions may be considered, as this pool of items may not be sensitive enough to distinguish between subjects with high and low performance.

Step D. The results of DIF analysis by gender allow to see that the DIF contrast in item P25 (*hydrocarbons*) was 0.91 and statistically significant ($p < 0.01$) and an effect size of 1.07. The Item difficulty (DIF mean) for the male sample was 0.02 logits while that for women was 0.091. This indicates that this question is *more difficult* for women. In the DIF analysis according to age, it was observed a statistically significant contrast ($p < 0.01$) of 0.66 and 0.80, and an effect size was 0.80 and 0.96, respectively, in items P29 (*cell*) and P42 (*hydrocarbons*). Both items are more difficult for the group of young persons aged between 18 and 20.

Form B

Step A. The RMSR value (0.0151) is lower than the typical residual estimated error (0.227) and the CFI was 0.84, which indicates that the assumption of unidimensionality is true. The values of the covariance matrix do not exceed the cut-off value (0.25) and only 8.5% of the residual with values higher than 0.025 was observed.

Step B. All 60 items showed an adequate fit to the model (*Mnsq* ≤ 1.3 or ≥ 0.6). Seven items, however, showed inadequate fit to the model (*Zstd* ≤ ±2.0) and *rpbis* values that were negative or close to zero. A model with 53 items was estimated where a pattern of a satisfactory fit for both *Mnsq* and *Zstd* was observed. The difficulty index (δ_i) of the items varied between $-2.25 \leq \delta_i \leq 1.65$ (*M* = 0.00, *SD* = 0.90). The analysis of the persons' adjustment reflects that 92% of the response patterns adjusted to the model. The skill levels varied between $-1.83 \leq \theta \leq 2.14$, (*M* = 0.05, *SD* = 0.77).

Step C. The item separation (6.80) and item reliability (.98), as well as the person separation (2.24) and person reliability (.83) values were satisfactory.

Step D. According to the gender, a DIF contrast of 0.76 ($p < 0.01$) and an effect size of 0.84 in item P51 (*hydrocarbons*) was observed, this item resulting more difficult for women. Regarding age, a contrast of 0.70 ($p < 0.01$) and an effect size of 0.78 was observed in P56 item (*cell*), this item resulting more difficult for the group whose ages varied between 18 and 21 years.

Anchor Items

Step E. The results showed a value of $r = .961$. The constant value was 0.01 and $\beta = .970$, so we can assume the invariance of the parameters of the anchor items.

Items of Level II

Form A

Step A. The RMSR value (0.0135) is lower than the typical residual estimated error (0.231) and the CFI was 0.84, which indicates that the assumption of unidimensionality is true. The values of the covariance matrix do not exceed the cut-off value (0.25) and only 6.6% of the residual with values higher than 0.025 was observed.

Step B. All 60 items showed an adequate fit to the model (Mnsq ≤ 1.3 or ≥ 0.6). Eight items, however, showed inadequate fit to the model

(*Zstd* ≤ ±2.0) and *rpbis* values that were negative or close to zero. A model with 52 was estimated where a pattern of a satisfactory fit was observed. The difficulty index (δ_i) of the items varied between $-3.40 \leq \delta_i \leq 2.19$ ($M = 0.00$, $SD = 1.30$). The analysis of the persons' adjustment reflects that 79.4% of the response patterns adjusted to the model. The skill levels varied between $-2.23 \leq \theta \leq 2.05$, ($M = 0.18$, $SD = 0.81$).

Step C. The item separation (8.84) and item reliability (0.99), as well as the person separation (2.18) and person reliability (0.83) values were satisfactory.

Step D. According to gender, no significant DIF contrast at $p < 0.01$ was observed. However, four items showed a statistically significant DIF contrast for age ($p < 0.05$). Two of these questions are more difficult for the group of young persons between 18 and 21, and two of them more difficult for the group of young adults between 22 and 48 years. However, the effect size was small (between 0.54 to 0.74.), and bias against boys relative to girls for items could explained by chance.

Form B

Step A. The RMSR value (0.0148) is lower than the typical residual estimated error (0.230) and the CFI was 0.86. The values of the variance and covariance matrices do not exceed the cut-off value (0.25) and only 8.5% of the residual with values higher than 0.025 was observed.

Step B. All 60 items showed an adequate fit to the model (*Mnsq* ≤1.3 or ≥ 0.6). Eight items showed inadequate fit to the model (*Zstd* ≤ ±2.0) and *rpbis* values and were eliminated A model with 52 was estimated where a pattern of a satisfactory fit was observed. The difficulty index (δ_i) of the items varied between $-3.56 \leq \delta_i \leq 1.57$ ($M = 0.00$, $SD = 1.16$). The analysis of the persons' adjustment reflects that 85.8% of the response patterns adjusted to the model. The skill levels varied between $-3.07 \leq \theta \leq 1.89$, ($M = -0.32$, $SD = 0.91$).

Step C. The item separation (8.08) and item reliability (0.98), as well as the person separation (2.60) and person reliability (0.87) values were satisfactory.

Step D. According to gender, a DIF contrast of 0.91 ($p < 0.01$) and an effect size of 0.78 in item P35 (*organ systems*) was observed, this item resulting more difficult for women. Regarding age, four items presented a DIF contrast statistically significant ($p < 0.01$) and the effects size was between 0.54 to 0.74. One of these questions is more difficult for the group of young persons between 18 and 21, and three of them are more difficult for the group of young adults between 22 and 48 years.

Anchor Items

Step E. The results showed a value of $r = .994$. The constant value was 0.02 and $\beta = .994$, so we can assume the invariance of the parameters of the anchor items.

Discussion

This study aims to develop a bank of items to measure the level of knowledge on biology using the Rasch model (Rasch, 1960). This psychometric model ensures that the parameters of persons and items be expressed in the same units (joint measurement), adjusts the data to the model showing which persons are independent from the administered items (specific objectivity), and ensures that the scale present interval properties (properties of measurement), such as the *logit* type (Fernandez, 2010). A calibrated IB is the prerequisite for CAT applications and can be used to generate tests on paper of fixed length measuring specific levels of knowledge required by teachers.

Overall, the items that make up the IB have acceptable psychometric properties. The difficulty and skill level indexes of the evaluated persons covered much of the measured continuum and the reliability indexes (persons and items) indicate that the location of the persons and items would be predictably reproducible (Andrich, 2002). The overall fit of the items was adequate, although a small number of them presented a differential

behavior according to the students' sex or age. In addition, it was observed that between 79.4% and 96% of the participants responded consistently to the test items, which allowed to identify patterns of predictable responses by the proposed model (Linacre, 2002). The study of invariance confirms that the parameters obtained from the anchor items in two subsamples are similar.

The results achieved are encouraging. Among the benefits offered by this IB is the adaptation of the curriculum to the requirements and needs of the students (Fuentes Navarro, 2004). That is, teaching would benefit if the contents and the difficulty of the instruction matched with the knowledge and skills of the subject, which would optimize the teaching process (Rolfhus and Ackerman, 1999). Furthermore, this assessment would allow to evaluate the quality of the educators' instruction (Popham, 2001). Finally, the evaluation may also be due to the criteria of quality control and efficiency of educational policies adopted (Martínez Rizo, 2009). Thus, having a measurement tool that has been properly developed would represent an advance in the assessment of educational systems learning (Arias, 2006).

On the other hand, these items can be loaded into specialized software and thus computerized adaptive tests (CAT) can be used, which would lead to minimize the standard error of measurement and the possibility of measuring length without loss of accuracy and reliability (Ibáñez, Arce, and Pareja, 1999). This would improve the diagnostic possibility with shorter and more accurate assessments (Olea and Ponsoda, 2003). This would help to follow a longitudinal track of an individual student's knowledge, to generate a diagnosis of the quantity and quality of acquired content, to specify which taught theoretical content is more difficult, and to incorporate new learning alternatives.

The contribution of this work is significant in the field of educational assessment in Argentina. However, it is relevant to mention some limitations. First, the sample size was sufficient to give some partial results and not results from all the forms of all levels. From the pool of 487 items, only 180 were analyzed. This is because the registration of students in the early years (first and second) is higher than in the last years of the majors (third and fourth). In the coming years, it is expected to have enough participants to perform calibrations of these items. Another limitation was related to the fact that no evidence of validity was generated with an external variable, as it could be a validity study of convergence. Future research could compare the performance in the TCG-B and other external criterion, such as students' academic performance.

By considering these limitations and the results achieved, it is expected soon to use a Biology Item Bank as a measure that allows to establish the advancement of student learning. In this way, the institutions involved (Faculties of Biology or Medicine, for example) might provide a report on the knowledge level achieved by each student, highlighting the strengths and weaknesses that need to be reviewed, as well as possible training courses that could allow a greater specialization in this field. This material would be a fruitful evidence of validity of the same test.

In short, the resulting pool of items allows to estimate the level of knowledge of the students who have attended university courses related to natural science and health science. This set of items will be part of a General Knowledge Item Bank that shall consist of twenty specific knowledge domains (Argentine Literature, for example) grouped into four domains of more general knowledge: Humanities, Science, Civic Education, and Mechanics (Cupani, Zalazar Jaime, Garrido, Gross, and Tavella, 2012; Cupani, Zalazar-Jaime, Ghío and Castro Zamparella, 2013). It is estimated that this item bank will comprise approximately 10,000 questions to measure these 20 specific domains (about 500 items per domain), distributed in four levels of instruction. For its calibration, a sample of about between 60,000 and 90,000 university students is needed.

Acknowledgement

This work was supported by grants to MC from the National Secretariat of Science and Technology (FONCYT) and from the Secretariat of Science and Technology - National University of Córdoba (SECyT-UNC).

References

Andrich, D. (2002). Understanding Rasch measurement: Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation. *Journal of Applied Measurement*, *3*, 325-359.

Arias, F. (2006). *El Proyecto de Investigación: Introducción a la Metodología Científica*. 5ta Edición. Caracas: Episteme.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Barrada, J. R., Olea, J., Ponsoda, V., and Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement*, *34*,(6), 438-452.

Bond, T. G. (2003). Relationships between cognitive development and school achievement: A Rasch measurement approach. *On the Forefront of Educational Psychology*, 37-46.

Bond, T. G., and Fox, C. M. (2007). *Fundamental measurement in the human sciences*. Chicago, IL: Institute for Objective Measurement.

Campbell, N. A. (2007). *Biología*. Séptima edición. Ed. Médica Panamericana.

Cupani, M., Zalazar Jaime, M. F., Ghío, F., and Castro Zamparella, T. (2013) Construcción de Distractores. Congreso; XXXIV Congreso Interamericano de Psicología. Brasilia.

Cupani, M., Zalazar-Jaime, M. F., Garrido, S, Gross, M., and Tavella, J. (2012) Construcción de un test de conocimiento general. X Congreso Latinoamericano de Sociedades de Estadística Córdoba, Argentina.

Downing, S. M., and Haladyna, T. M. (2006) *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Eggen, T. J. (2011). Computerized classification testing with the Rasch model. *Educational research and evaluation*, *17*(5), 361-371.

National Research Council. (2009). *Learning science in informal environments: People, places, and pursuits*. National Academies Press.

Fernández, J. M. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del psicólogo*, *31*(1), 57-66.

Fraser, C., and McDonald, R. P. (2012). NOHARM 4. Manual. Retrieved on May 24, 2014, from http://noharm.niagararesearch.ca/nh4man/nhman.html

Fuentes Navarro, R. (2004). La constitución científica del campo académico de la comunicación en México y en Brasil: análisis comparativo. Protocolo de investigación presentado al Comité de Ciencias Sociales del Fondo de Ciencia Básica del Consejo Nacional de Ciencia y Tecnología.

Garbanzo Vargas, G. M. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación*, *31*(1), 43-63.

Gauggel, S., Heinemann, A. W., Böcker, M., Lämmler, G., Borchelt, M., and Steinhagen-Thiessen, E. (2004). Patient-staff agreement on Barthel index scores at admission and discharge in a sample of elderly stroke patients. *Rehabilitation Psychology*, *49*(1), 21-27.

Haladyna, T. M., Downing, S. M., and Rodríguez, M. C. (2002). A review of multiple-choice item writing guidelines. *Applied Measurement in Education*, *15*()3, 309-334.

Haladyna, T. M., and Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied measurement in education*, *2*(1), 51-78.

Hambleton, R. K. (2000). Response to Hays et al and McHorney and Cohen: Emergence of item response modeling in instrument development and data analysis. *Medical Care*, *38*(9), II-60.

Hambleton, R. K., and Swaminathan, H. (1985*). Item response theory: Principles and applica-*

tions (Vol. 7). Springer Science and Business Media.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). Fundamentals of item response theory (Measurement methods for the social sciences series, Vol. 2).

Lai, J. S., Dineen, K., Reeve, B. B., Von Roenn, J., Shervin, D., McGuire, M. et al. (2005). An item response theory-based pain item bank can enhance measurement precision. Journal of pain and symptom management, *30*(3), 278-288.

Fernández, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85-106.

Linacre, J. M., and Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*.

Linacre, J. M. (2009). Winsteps (Version 3.48) [Computer software and manual]. Chicago, IL:[Online]. Retrieved June 25.

Long, C., Wendt, H., and Dunne, T. (2011). Applying Rasch measurement in mathematics education research: Steps towards a triangulated investigation into proficiency in the multiplicative conceptual field. *Educational Research and Evaluation*, *17*(5), 387-407.

Martínez Rizo, F. (2009). Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado. *Revista electrónica de investigación educativa*, *11*(2), 1-18.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Novak, J. D., Mintzes, J. J., and Wandersee, J. H. (2000). Learning, teaching, and assessment: A human constructivist perspective. Assessing science understanding: A human constructivist view, 1-13.

Ibáñez, J. G. M., Arce, J. S., and Pareja, I. (1999). Desarrollo de un sistema informático orientado a la construcción y gestión de bancos de ítems. In *Tests informatizados: Fundamentos y aplicaciones* (pp. 85-108).

Olea, J., and Ponsoda, V. (2003). Test adaptativos informatizados. Madrid. UNED.

Popham W. J. (2001). The truth about testing: An educator's call to action. Alexandria, VA: Association for Supervision and Curriculum Development.

Prieto, G., and Delgado, A.R. (2003). Análisis de um test mediante el modelo de Rasch. *Psicothema*, *15*(1), 94-100.

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

Rolfhus, E. L., and Ackerman, P. L. (1999). Assessing individual differences in knowledge: Knowledge, intelligence, and related traits. *Journal of Educational Psychology*, *91*(3), 511.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*(3), 159-203.

Törmäkangas, K. (2011). Advantages of the Rasch measurement model in analysing educational tests: An applicator's reflection. *Educational Research and Evaluation*, *17*,(5), 307-320.

Wright, B. D., and Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement*, *21*,(4), 331-345.

Wright, B. D., and Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*(12), 857-860.

Wright, B. D., and Stone, M. H. (2003). Five steps to science: Observing, scoring, measuring, analyzing, and applying. *Rasch Measurement Transactions*, *17*, 912-913.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-214.