Universidad de Costa Rica
Facultad de Educación
Instituto de Investigación en Educación

REVISTA ACTUALIDADES INVESTIGATIVAS EN EDUCACIÓN

http://revista.inie.ucr.ac.cr/ ISSN 1409-4703

DIF in Spanish and Mathematics from Costa Rica's national tests in reported students with ADHD

Funcionamiento diferencial del ítem en pruebas de español y matemática en estudiantes reportados con TDAH

Volumen 17, Número 2
Mayo-Agosto
pp. 1-22

Este número se publica el 1° de mayo de 2017
DOI: http://dx.doi.org/10.15517/aie.v17i1.28661

Eiliana Montero-Rojas
Tania Elena Moreira-Mora

Revista indizada en **REDALYC**, **SCIELO**

Revista distribuida en las bases de datos:

LATINDEX, **DOAJ**, **REDIB**, **IRESIE**, **CLASE**, **DIALNET**, **SHERPA/ROMEO**, **QUALIS-CAPES**, **MIAR**

Revista registrada en los directorios:

ULRICH'S, **REDIE**, **RINACE**, **OEI**, **MAESTROTECA**, **PREAL**, **CLACSO**

DIF in Spanish and Mathematics from Costa Rica's national tests in reported students with ADHD

Funcionamiento diferencial del ítem en pruebas de español y matemática en estudiantes reportados con TDAH

Eiliana Montero-Rojas[1]
Tania Elena Moreira-Mora[2]

Abstract: *The detection of differential item functioning (DIF) is fundamental to ensure instruments' invariance, and, therefore, a better estimate of the construct being measured across the different groups of examinees. The purpose of this research was to provide substantive hypothesizes related to possible sources of DIF, comparing students reported with accommodations for ADHD (focal group) and students with no accommodations (reference group), using the Standardized-P Difference and the Mantel Haenszel (MH) methods. Data from the Costa Rican national high school exit tests in Spanish and Math, from the year 2004, in public schools, were analyzed. First, these two methods were used to detect items with DIF, and then, using a more qualitative approach, drew hypotheses related to possible sources of DIF. Some degree of agreement was verified between the two different empirical methods, being Mantel-Haenszel more sensitive. In the Spanish test, DIF was hypothesized to be caused by the length and wording of the options, and the literary and non-literary texts in the stem. In Math, possible sources of DIF involved vocabulary, wording, the transition of verbal to mathematical language, the visuospatial item organization, and the drawing of graphs and geometrical figures. At the methodological level, complementing the statistical analyses with the judges' criteria was helpful to identify possible sources of irrelevant variance in the construct measured by these tests. The hypotheses must be interpreted with caution, though, since the number of items detected as exhibiting DIF was relatively small.*

Key words*: DIF, validity, testing accommodations, adhd, mathematics, language.*

Resumen: *La detección del funcionamiento diferencial del ítem (FDI) es fundamental para garantizar instrumentos invariantes y una mejor estimación del constructo en los diferentes grupos de examinados. El propósito de esta investigación fue proveer hipótesis sustantivas sobre posibles fuentes de FDI, comparando estudiantes reportados con el trastorno de déficit atencional con hiperactividad (TDAH, grupo focal) y estudiantes sin esas características (grupo de referencia). Se usaron los métodos de la diferencia p estandarizada y Mantel-Haenszel para identificar los ítems con FDI en las pruebas de bachillerato de español y matemática aplicadas en el año 2004 en colegios públicos académicos de Costa Rica. Luego, usando un enfoque más cualitativo, se generaron hipótesis sobre sus posibles fuentes. Hubo cierto grado de concordancia entre ambos métodos, siendo el de Mantel-Haenszel más sensible. En la prueba de español se encontró evidencias para apoyar la hipótesis de que la extensión y la redacción de las opciones y los textos literarios y no literarios incluidos en el encabezado pueden ser causas de FDI. En matemática se encontró que el vocabulario, la redacción, la transición del lenguaje verbal al matemático, la organización viso espacial del ítem y el dibujo de gráficas o figuras geométricas pueden ser causas de FDI. A un nivel metodológico, el complementar los análisis estadísticos con el criterio de jueces fue útil para la identificación de posibles fuentes de variancia irrelevante al constructo medido por estas pruebas. Las hipótesis deben ser tomadas con cautela, ya que el número de ítems detectados con FDI fue relativamente pequeño.*

Palabras clave*: DIF*, *validez, adecuaciones, TDAH, matemática, español.*

[1] *Investigadora de la Universidad de Costa Rica, en el Instituto de Investigaciones Psicológicas. Dirección electrónica: eiliana.montero@ucr.ac.cr*

[2] *Investigadora del Instituto Tecnológico de Costa Rica. Dirección electrónica: tmoreira@itcr.ac.cr*

Artículo recibido*: 9 de marzo, 2016*
Enviado a corrección*: 7 de noviembre, 2016*
Aprobado*: 18 de abril, 2017*

1. Introduction

This investigation studies DIF for items in Math and Spanish exams from the Costa Rican national exit tests in High School. DIF has currently become one of the key issues in test validation and takes into account the wide use of tests for selection, promotion and certification purposes in education (Hidalgo, Galindo, Inglés, Campoy y Ortiz, 1999). As authors Gómez-Benito, Hidalgo and Guilera (2010) point out, DIF, as any aspect of validity, involves a process to accumulate evidences. However, research related to identifying DIF or its sources has not been carried out systematically in Costa Rica for its national standardized tests. Moreover, DIF research for the population with ADHD (Attention Deficit Hyperactivity Disorder) had not been undertaken at the time of the present study. No published studies were located around this topic with data from the Costa Rican population. The admission test program at the University of Costa Rica, who develops the oldest standardized test in Costa Rica, with more published scientific articles than any other testing program in the country, has only calculated DIF regularly for more traditional comparisons, gender and high school status (public and private) (Montero, personal communication, 20th Octuber 2016).

In general, international studies have focused in minority groups' performance in Math and Language, and the effects on DIF provoked by cultural variables and linguistic traits of the items in the tests (Abedi, Hofstetter, Baker and Lord, 2001). There are also studies regarding the effects of test accommodations in the tests' scores of students' with disabilities, (Thompson, Blount y Thurlow, 2002). Other research pertains to the effect of accommodations in the test's validity for students with disabilities (Koretz y Barton, 2003). In Costa Rica, this is the first DIF study carried out with the ADHD population in a national standardized test. Its importance is precisely to provide evidence of test validity for this specific population. This issue is paramount since these tests are of high stakes for the students.

The general objective this research is to evaluate DIF in items taken from the High School Education Exit Tests in Spanish and Math in Costa Rica, comparing students with and without accommodations for ADHD (Attention Deficit Hyperactivity Disorder) from public day schools in the year 2004. Moreover, the specifics objectives are:

a. To identify DIF in items from the Spanish and Math tests applying the empirical methods of the Standardized P Difference and Mantel-Haenszel.

b. To determine possible sources of DIF in items from the Spanish and Math tests comparing students with and without accommodations for ADHD from public day schools in 2004.
c. To generate theoretical hypotheses to explain and control, a priori, DIF in these exams for students with ADHD.

Within this context, this investigation aims at contributing, on one hand, to provide a methodological approach for a more comprehensive study of DIF and, on the other hand, to generate various theoretical hypotheses about sources of DIF for the population of students with ADHD. These hypotheses are also useful to control DIF a priori in future tests of Spanish (as a native language) and Math, for this testing program.

2. Theoretical Framework

2.1 Attention Deficit Hyperactivity Disorder

In the Costa Rican context, Special Education is the set of support services that are available for the students with special educational needs, whether temporarily or permanently (MEP, 1998, p.5). This is an issue of interest for this study, since the population analyzed is made up of two subgroups: the reference group (RG), i.e. students that do not report special needs, and the focal group (FC) who are students registered with behaviors that indicate the presence of ADHD. According to Act 7600 that provides for equality of opportunities for people with disabilities and that has been in effect since May 1996, students with special needs are integrated to the regular school system with the support and curricular accommodations necessary to guarantee equal and quality educational opportunities as compared to regular students. Consequently, the purpose of accommodations is to meet the special educational needs of the students reported with some type of disability, since, arguably, their deficit brings out a higher degree of difficulty to learn, compared to their peers.

In accordance with access policies, a curricular accommodation involves the adjustment of the educational offer to the characteristics and needs of the student, with the purpose of catering to the individual differences. These access accommodations can be significant or not. This study used a focal group comprised of students with non-significant accommodations, consisting of an additional hour for answering the test, and, the presence of a specialized tutor during the test administration.

The current scientific literature around the topic of ADHD agrees in pointing out that its primary deficit is associated with an executive dysfunction, According to Rubiales, Bakker, Russo and González (2016) these executive functions are defined as a set of cognitive abilities that allow the individual to establish objectives, planning, initiating activities, task monitoring, selecting behaviors and carrying out actions to achieve the target objectives, showing a behavior that is effective, creative and socially accepted. ADHD is one the disorders affecting the neurobiological development dysfunctions of the child. In psychological terms, ADHD comprises "a heterogeneous group of clinical manifestations whose most visible behavioral materializations are, broadly speaking, hyperactivity, impulsivity, and the difficulty to maintain attention" (Moreno, 2001, p. 81). According to this author, ADHD involves a set of visible behavioral traits that contribute to the inadaptability of children and young people to their academic and family environments. This group is usually characterized by the inability to remain focused on every day and academic activities, according on three specific areas: concentration capacity, impulse control and, in some cases, activity levels (Villalobos and Morales, 2002).

2.2 Validity evidence

Regarding the characteristics of the Math and Spanish tests administered to both groups of students (with test accommodations for ADHD and, with no test accommodations), the model for the interpretation of scores is norm referenced, based on the normal curve and the discrimination between students in terms of their relative level of learning in these subject matters. The purpose of the tests is certification. The comparison of scores obtained by the examinees depends on the psychometric properties of tests in this particular application and purpose, particularly unidimensionality, reliability and validity.

This study understands validity as an integrated and evaluative judgment of the degree in which the empirical evidence and the theoretical reasons support the adequacy and appropriateness of the interpretations and actions based on tests scores or other assessments tools (Messick, 1995, p. 5). Thus, in accordance with the standards for educational and psychological tests established by the American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME), the validation process involves the accumulation of evidence that provides a scientific base for the interpretation and relevance of test scores (AERA, APA y NCME, 2014).

The essence of this conceptual perspective is the integration of the three types of traditional evidence: content, criterion and construct, by the unifying thread of construct validity, and thus supporting the adequacy of the interpretations and uses of test scores, including the social consequences derived from these interpretations. Messick (1989) establishes two interrelated aspects: the source of the justification of the test, based on the study of the evidence that supports the meaning of the score, and secondly the function or the result of the test, i.e. interpretation and use. Within this unified perspective, DIF analysis is instrumental for providing empirical evidence of the degree in which the scores properly measure a particular construct.

2.3 Differential Item Functioning

This analysis has been frequently confused with bias, especially by the twofold meaning of the latter (the social and the statistical meaning). In this respect, Angoff (1993) has pointed out that this has generated an unnecessary confusion, given that some use the term to describe the judgment or assessment of bias with a social perspective while others refer to the statistical observations. Actually, "DIF" is employed for the statistical properties of the item in different groups (Prieto, 2013).

From a psychometric perspective, an item shows differential functioning if subjects with an identical level in the trait being measured and belonging to different subpopulations or cultural groups do not have the same probability of correctly answering the item. (Anastasi and Urbina, 1998; Attorresi, Galibert, Zanelli, Lozzia and Aguerri, 2003; Camilli, 1993; Hidalgo et al., 1999; Muñiz, 1990; Padilla, González and Pérez, 1998; Penfield and Camilli, 2006). The comparison is usually carried out between the group of main interest, called the focal group, and the group that serves as the basis of comparison, the reference group (Donoghue, Holland and Thayer, 1993; Hidalgo et al., 1999; Montero, 1993). As a result, an item presents DIF if, under equal conditions, examinees belonging to the reference group systematically show a different probability of answering the item correctly compared with examinees from the focal group.

In the sixties, multiple and rigorous statistical procedures arose for the detection of DIF that were classified in two broad categories (Bandeira, 2002, 2003; Hidalgo, López and Sánchez, 1997; Montero, 1993; Wainer, 1993). The first includes the empirical methods (also known as observed conditional invariance methods or conditional methods) based on the observed scores in the test, from the perspective of the Classical Test Theory (CTT). The

second includes the theoretical methods (also known as non-observed conditional invariance methods or unconditional methods) grounded in mathematical models such as Item Response Theory (IRT), based on estimations of ability (θ), derived from the model that is more appropriate for the data (Gómez-Benito et al., 2010).

In the current study, two empirical methods were selected to identify the items exhibiting DIF, Mantel-Haenszel and the Standardized P Difference. The reason that justifies this selection has to do with the measurement model that provides the frame of reference for the construction of the Math and Spanish tests, which is CTT (Classical Test Theory). According to Hidalgo et al. (1997) the selection of the method to detect DIF depends on the characteristics of the measurement model. Thus, if the test has been developed under the CTT, empirical methods should be used. Even though these two methods are not new, and there are currently new proposals for DIF detection, they are still widely used, as it is shown in a recent publication by Gómez-Benito, Balluerka, González, Widaman and Padilla (2017).

The Mantel-Haenszel method provides an estimate of the DIF magnitude called the Mantel-Haenszel common odds ratio estimate MH (αMH), and a test for its statistical significance known as Chi-squared MH ((χ2 MH), with one degree of freedom. This method is based on the comparison of observed and expected frequencies of number of right and wrong answers in an item, by subjects that, belonging to different populations (focal and reference groups) display the same level in the test total score (Bandeira, 2002, 2003; Elosúa and López, 1999; Hidalgo et al., 1999; Longford, Holland and Thayer, 1993). In these calculations, the latent variable θ, represented by the total score, is divided in K ability intervals; K 2x2 contingency tables are built for each item. Subjects are classified according to group membership (focal or reference) and the possible item responses to the item in these tables (Bandeira, 2002, 2003; Hidalgo et al., 1997, 1999; Longford et al., 1993).

The odds-ratio (Θ_{M-H}), expresses the ratio between the probability of correctly answering the item against the probability of failing it in the focal group, and the probability of answering it correctly against the probability of getting it wrong in the reference group (Bandeira, 2002; Hidalgo et al., 1997, 1999). The Mantel-Haenszel quotient of ratios is obtained by the following expression:

$$\Theta_{M-H} = \frac{\Sigma A_k D_k / T_k)}{\Sigma C_k B_k / T_k}$$

Where k (lowercase) is the specific kth score level, being K (capital letter) the total number of score categories. Ak represents the numbers of examinees in the reference group, at the kth score level, who answered correctly the item, where B includes those examinees from the reference group that answered incorrectly the item. On the other hand, Ck is the number of examinees in the focal group that answered correctly, and Dk is the number of people in that group that answered incorrectly the specific item at the kth score level. Tk is the total number of examinees in that score level. The sum includes all score levels.

The value of $\Theta_{M\text{-}H}$ can range between 0 and ∞. If $\Theta_{M\text{-}H}$ is higher than 1, it favors the reference group and, if lower, it means the focal group displays a higher performance when compared with the reference group (Bandeira, 2002). However, because of practical reasons, Penfield (2013), in his software package DIFAS proposes a transformation of the odds ratio with an asymptotic normal distribution and a logarithmic behavior. This is the approach used in the present study. If odds ratio > 1 then the log odds ratio is positive, indicating DIF in favor of the reference group, and, if 0 < odds ratio < 1, then the log odds ratio is negative, implying DIF in favor of the focal group. In the case of dichotomous items, DIFAS also provides a classification of the coefficients based on criteria by the Educational Testing Service (Carvajal and Poggio, 2006).

According Carvajal and Poggio (2006) the Educational Testing Service (ETS) has proposed the use of a hierarchical scale for the different values of the ∆MH coefficient (logarithm of the odds ratio in a delta metric), also known as "Mantel Haenszel delta difference" (MH D-DIF), depending on its magnitude. This classification in based on two factors, the absolute value of MH D-DIF, and, whether or not this value shows statistical significance at a probability level of .05 (p =.05). Both factors must be taken into account, not only the statistical significance, since there are cases with very small magnitude of MH D-DIF, but statistically significant, given that the analysis used a very large number of examinees (Zieky, 1993). The three categories of the MH D-DIF have been labeled with A, B and C (Bandeira, 2002; Dorans and Holland, 1993; Longford et al., 1993; Prieto, 2013; Zieky, 1993).

The category A items include those with non-statistically significant values of MH D-DIF (p > 0,05) or with less than 1 absolute values of MH D-DIF (delta unit). DIF for items in this category is considered negligible or insignificant. Items in the C category are in the other extreme, and they are considered to have strong evidence of DIF, these are items with statistically significant values of MH D-DIF (p < 0,05) and with absolute values of MH D-DIF (delta unit) equal or greater than 1,5. The middle category, B, includes items that don't match

the definition to be included in categories A or C. For example, an item with a statistically significant MH D-DIF, but with a smaller than 1,5 absolute value will be included in this category, also an item with a greater than 1,5 MH D-DIF, but not statistically significant.

Another empirical model that has been widely used for identifying DIF is the Standardized P Difference (STD P-DIF). In this case, an item will exhibit DIF when the expected performance for individuals with the same degree of ability, but belonging to different groups, is dissimilar (Dorans and Holland, 1993; Montero, 1993). With this method, a discrepancy index is calculated between the two groups regarding the performance in the item (p difference) based on the expression described by Montero (1993):

$$\Sigma [Ks (pfs – pbs)] / \Sigma Ks \qquad (2)$$

Where:

pfs = is the rate of right answers in the focal group (minority) at the "s" ability level.

pbs = is the rate of right answers in the base or reference group (majority) at the ability level "s".

Ks = is the weight factor for each level of the "s" score, i.e., the number of subjects in the focal group at the "s" level.

The STD P-DIF is an index that can take values between -1 y 1 (or -100 and 100), and its direction is provided by the + or - sign. Positive values indicate that the item favors the focal group whereas the negative values indicate DIF against the focal group. Values of the STD P-DIF are organized in a hierarchical scale, proposed by ETS, according to its magnitude (Bandeira, 2003; Dorans and Holland, 1993; Montero, 1993; Prieto, 2013). The use of both empirical methods in this research also fulfills the need to follow the current trend to apply two or more procedures to detect DIF.

3. Method

This research is non-experimental and uses mixed methods to fulfill its goals. It's non-experimental since it was limited to the observation of DIF in the tests, without introducing any alteration in the educational treatment, in the administration and construction of the exams, or in the students' scores. Given that it is focused on a research problem that has never been studied in the Costa Rican educational context before, it is classified as exploratory.

According to Cea (1999), as far as the temporal frame is concerned, it is a cross-sectional design, since the data were collected from the tests applied in one single administration in 2004.

3.1 Sources of information

The study was carried out with the population of students who took the national exit tests of secondary education in Spanish and Math, in 2004. It included 217 public high schools. There were 14510 students who took the Spanish test and, 15042 who took the Math test. From those 476 students in Spanish and 493 in Math had test accommodations for ADHD. These accommodations represent around 82,5% of the total population with special needs receiving accommodations that year. The students with accommodations for ADHD were the focal group. The reference group included students with no accommodations, these were 13659 in Spanish and 14132 in Math. In DIF analysis the studied groups are identified as Reference Group (RG) and Focal Group (FG) (Aguerri, Blum, Picón y Galibert, 2010). The reference group had no accommodations at all, and most of the accommodations for students with ADHD involved more time for taking the tests. All data were treated under conditions of anonymity and the researchers had the approval of Ministry of Public Education.

The other sources of information for this study were teachers in Math and Spanish and professionals in Special Education who participated as judges to determine possible sources for DIF. In addition, six students from the same population with ADHD accommodations were studied using observation and discussion group techniques.

3.2 Procedure and techniques

The procedure was carried out in two main stages: The first one was exploratory and started with a *"wandering"* phase, where two qualitative techniques were employed: participant observation of six students with ADHD accommodations in a group of last grade from a public high school, followed by a discussion group with two of these six.

The goal of the participant observation was first to grasp the manifestations of ADHD in the everyday activities of the youngsters, and, secondly, to generate some theoretical hypotheses about DIF for this particular population, observing their verbal and non-verbal communication, behaviors and inter-subjective relationships in the Spanish and Math lessons. This process involved twelve sampling sessions, six in Spanish and six in Math. A registry of the observation was written in a logbook-like format. Notes included descriptive information

and reflections about the students with ADHD's behaviors and interactions, verbal and non-verbal; they also included comments provided by the two teachers.

The discussion group technique consisted of gathering the six students in only one session. However only two of them attended this session, they were precisely the students who had more evident behavioral manifestations of ADHD symptoms, hyperactivity and impulsivity. The purpose of this activity was to go deeper in the students' perceptions about the difficulty level of the items. In this session the students identified first which items they perceived as easier or more difficult. Then, with an open ended questionnaire, the students were motivated to express their comments about the difficulty level of the items and possible reasons that explained that difficulty, in their particular cases.

The application of both qualitative techniques responded to an interest to understand the behaviors related to ADHD in their everyday life and to generate some theoretical hypotheses to be tested using the empirical evidence provided by the DIF analyses.

Then item analyses with Classical Test Theory were performed using Cronbach's Alpha to determine the degree of internal consistency of the Math and Spanish tests. The factor structure was also studied using exploratory factor analysis to obtain evidence regarding the unidimensionality in both tests. Unidimensionality is an assumption in Classical Test Theory, the measurement model from which these DIF detection methods are derived, therefore, it is important to explore the plausibility of that an assumption with the data one is working with. Finally, in this first stage, the empirical detection of DIF was carried out using the Standardized P Difference and the Mantel Haenszel.

The second stage involved an interpretative approach focused on the inquiry of possible sources of DIF using expert judges (Montero, 1993). Two groups of judges were formed, one was composed by three Spanish teachers with teaching experience at the grade level of the tests, and who had taught students with ADHD, along with two Especial Education specialists. The second team of judges was comprised by three Math teachers, with similar background characteristics to the teachers in the first group, and two other Especial Education specialists. First, they looked at different items, some with DIF and some without it, unaware of their DIF classification, and made a prediction regarding whether or not the item had DIF, and, if so, they had to say what the direction was (in favor or against the focal group). At the same time, they were asked to provide explanations related to possible irrelevant attributes in the items they suspected as causing statistically substantial DIF in both tests.

The degree of concordance between the statistical results and the judges' classification of the items in three categories (-1 favoring the focal group, 0 no DIF, and, 1 against the reference group) was estimated using Kendall's tau-b. In general, there was a moderate to negligible relationship between the judges' classifications and the statistical results. Two out of the five Spanish judges showed measures greater than 0.2 in tau-b (one was 0.56 and the other 0.27). Three of the Math judges showed tau-b values greater than 0.2 (0.27, 0.23 and 0.21). Only the explanations given by judges whose predictions were in agreement with the statistical results were used to draft and discuss hypotheses about possible sources of DIF in the analyzed items.

With this methodological approach, grounded in the hypothetical deductive method, and using a triangulation strategy with theory about the nature of ADHD, the statistical evidence and the qualitative data allowed us to achieve the objectives of the study.

4. Results

The main results of the research are summarized as follows, in the same order as the study objectives.

4.1 Detecting DIF

The Mantel Haenszel method was more sensitive for identifying DIF; it is also more comprehensive than the STD P DIF, since it includes a statistical significance test, based on Chi-Squared. Mantel Haenszel also detected all the items detected with the STD P DIF. In total, Mantel Haenszel detected 12.5% of the 48 items analyzed in the Spanish test with DIF against the focal group. According to the hierarchical scale established by ETS only one of these items resulted in having moderate DIF and the other five were classified in the A category (insignificant). These results are shown in the table 1.

Table 1 - *Summary of the differential functioning analysis of the items in the Spanish and Math test*

Item	Mantel Haenszel [1/]			Standardized p Difference	
	χ^2_{MH}	Log odds ratio	ETS	Discrepancy indicator	ETS
Spanish Test					
1	13,776	0,552	B		
4	4,107	0,211	A		
10	7,047	0,318	A		
17	6,979	0,261	A	-0,059	A
23	5,832	0,252	A	-0,053	A
59	5,627	0,276	A		
Math Test					
9	4,5238	0,2149	A		
10	3,8545	0,1939	A		
16	8,7844	-0,2943	A	0,064	B
18	5,1199	-0,4306	A		
20	6,7964	0,3339	A		
26	6,7687	0,3052	A	-0,052	A
34	4,3479	0,267	A		
37	4,4699	-0,2129	A		
41	8,8674	0,3887	A	-0,055	A
45	8,035	0,3561	A	-0,054	A

Source: The data were provided by Ministry of Public Education
1/ A significance level of 0,05 was used for the hypothesis test of the Chi-Squared statistic in Mantel Haenszel

In the Math test, 17.9% of the 56 items analyzed were detected as presenting DIF; of those, 5.4% was favorable to the focal group (3 items) and 12.5% to the reference group (7 items). According to the ETS scale, 9 items were classified in the A category and just one with a moderate magnitude (B category).

As it was mentioned before, these results show higher sensitivity of Mantel Haenszel over the STD PD to detect DIF in these samples. Perhaps one of the reasons for this behavior is the relatively large sample sizes for this particular study.

Moreover, the mean differences between both groups were statistically significant, according to the t test for comparison of two independent samples means. This happened in both tests, Spanish and Math, considering all the items in the test (56 in Math and 48 in Spanish), and considering only the mean of the items with no DIF (46 in Math y 42 in Spanish). These results are shown in the table 2.

Table 2 - ***Descriptive Statistics for the items in reference and focal groups for the Spanish and Math tests***

	Group	N	Mean	Standard deviation	Mean standard error
Mathematics					
Total_56	Focal	493	23.96	8.920	.402
	Reference	14132	27.40	8.442	.071
Only Items with no DIF_46	Focal	493	20.47	7.652	.345
	Reference	14132	23.26	7.227	.061
Spanish					
Total_48	Focal	476	27.31	6.121	.281
	Reference	13659	31.51	6.263	.054
Only Items with no DIF_42	Focal	476	23.60	5.566	.255
	Reference	13659	27.14	5.698	.049

Source: The data were provided by Ministry of Public Education

The degree of coincidence between the statistical results and the judges' classification of the items in three categories (-1 favoring the focal group, 0 no DIF, and, 1 against the reference group) was estimated using Kendall's tau-b. In general, there was a moderate to negligible relationship between the judges' classifications and the statistical results. Two out of the five Spanish judges showed measures greater than 0.2 in tau-b (one was 0.56 and the other 0.27). Three of the Math judges showed tau-b values greater than 0.2 (0.27, 0.23 and 0.21). Only the explanations given by judges who were in agreement with the statistical results were used to draft and discuss hypotheses about possible sources of DIF in the analyzed items.

4.2 Sources of DIF

According to the judges' consensual criterion, there were mainly two aspects of the item structure that possibly explain DIF in the Spanish test: the first one has to do with the length and wording of the options, and the second one with the semantic ambiguity of the literary texts and complex content of non-literary texts. Therefore, they turn out to be more difficult for students with ADHD compared with students with no ADHD at the same ability level. In other words, the plurality of meanings of the literary text, in conjunction with the length, structural complexity and content difficulty of some of these items make them more difficult for ADHD students. In a particular item, for example, the student has to read 4 short fragments

extracted from literary works, and has to identify the fragment that uses a direct narrative style. The experts hypothesized as possible sources of DIF in this item the following irrelevant characteristics:

- Recognizing the narrative style in four texts is tiresome and confusing for many students, and even more so for those displaying ADHD behaviors, as they have to go back and forth to read the fragments.
- Including a variety of texts in the item increases item complexity; therefore, students with ADHD are not able to focus their attention on recognizing the narrative style due to the information overload in the four literary fragments.

The possible sources of DIF against the focal group in the Math items drafted by the judges include the use of inaccurate vocabulary, confusing writing, transition from verbal language to algebraic language, complex procedures due to the amount of computations, concepts and details, figure drawing, and spatial location of the mathematical expressions and figures. One case, for example, presents a complex structure, given the amount of concepts (secant, tangents, concentric), confusing vocabulary and having to draw a geometrical figure using the verbal information. This implies a conversion to mathematical language, thereby reducing the likelihood for focal group examinees to answer the item correctly.

The judges also made some conjectures, posteriori, related to possible reasons for DIF in the three items that favored the focal group, a result not expected. They pointed out the following item characteristics: the simple structure of the item, the use of accurate vocabulary, the measurement of just one Math concept and the short extension.

4.3 Discuss of theoretical hypothesis

As a result of the literature review, and integrating the knowledge from the data collected by observations and discussion group, a set of six hypotheses was generated a priori. The following paragraphs describe these hypotheses and their discussion with the results of empirical analyses.

4.3.1 Math hypothesis

Hypothesis 1: Math items that involve geometrical figures and graphs, adding multiple data in the stem, elicit DIF against students with ADHD.

Only item 48, presenting a geometrical figure in the stem, exhibits high DIF against the focal group. At this point it is relevant to point out that there was a general consensus among the five Math judges that the figures help the examinees to grasp the problem presented by the item, with the condition of not including unnecessary details that could confuse and mislead the students. In the case of this specific item, possibly, according to the judges, the complicated wording of the item in the stem, was the cause of DIF against the focal group, and not particularly the figure, that was specific and simple. This position is confirmed with the item #18 that presented DIF but in favor of the focal group, the graph provided by this item was considered specific and simple, as the one for item #48. On the other hand, there were two items that exhibited high DIF against the focal group, due to the fact that, possibly, the judges hypothesized, the students should draw the graph (#26) and the geometrical figure (#41).

Hypothesis 2: Math items that combine multiple data and concepts in the stem in order to make the student carry out different procedures and find the right answer contribute to high DIF against the students reported with ADHD behaviors.

The DIF analyses evidence was obtained to support this hypothesis, given that, from the 7 items with DIF against the focal group, 3 included multiple data and concepts in the stem. One of them, for example, portrays diverse data to calculate the length of a rectangle and solving an equation. Another item requires handling multiple concepts (concave function, vertex, intersection, decreasing function and quadratic function), drawing a graph and discriminating between two propositions. The third one combines multiple concepts such as: plane, distance, circumference, secant, concentric circles and radius to identify the circumference of the geometrical figure.

Hypothesis 3: Math items that include in the stem several propositions to be classified as true or false favor a high differential item functioning against students reported as presenting ADHD behaviors.

In this case, evidence was scarce, as only one item was detected. However, the judges mentioned that discriminating between two propositions is a characteristic that raises the complexity of the item, since it implies several calculations and applying reasoning skills in order to recognize the true proposition.

4.3.2 Spanish Hypothesis

Regarding the proposed theoretical hypotheses for the Spanish test in the current study, they mainly focused on item wording, length and structure as possible sources of high DIF against the group of examinees reported as exhibiting ADHD behaviors.

Hypothesis 4: Spanish items that include lengthy literary and non-literary texts in the stem contribute to a high differential item functioning against students reported as presenting ADHD behaviors.

For this hypothesis, evidence was gathered from the DIF analyses. For example, two items presenting the same format, i.e. four literary texts in the item stem, were detected with DIF against the focal group. This format, in the experts' point of view, increases the complexity of the item. Another irrelevant attribute was the length of the non-literary text.

Hypothesis 5: Spanish items that involve multiple concepts, events or statements to identify the right answer favor high DIF against students reported as exhibiting ADHD behaviors.

From the six items detected with high DIF against the focal group, judges recognized, among other possible sources, several statements that could be provoking DIF. Indeed, if we assess the clinical manifestations that characterize these young students with ADHD, tending to be very forgetful, they all have handicaps to organize their ideas, providing hasty or hurried answers and changing frequently their attention focus. One would expect that, when faced with this type of item, requiring a high level of abstraction or synthesis, their probabilities of answering correctly are lower, even though they do have the information to solve it.

Hypothesis 6: Spanish items with an open or unfinished text within the stem that has to be completed with the right answer, contribute to high DIF against students reported with ADHD behaviors.

It is important to highlight that 13 items in the Spanish test presented this format with an unfinished text in the stem, however only item #59 was detected to have a high DIF against the focal group. In the experts' point of view, probably the combination of the complexity of the concept with this item structure favored the presence of DIF. This result makes it impossible to substantiate this hypothesis. However, there was agreement among judges that this type of structure increases the complexity of the item, since completing the text with four possible answers affects examinees reported as having ADHD behaviors, as they require

more time to analyze each option and to identify the correct one, thereby creating more stress, anxiety and confusion.

Hypothesis 7: If there is a different factor configuration in the Spanish and Math tests applied to the students reported as having ADHD behaviors, compared to the students without special educational needs, that difference will be the DIF source.

The analysis of the factor structure of the Spanish and Math tests was carried out with the objective of identifying those items that measure a secondary component or factor in some of the groups compared, given that factorial difference would be a possible source of DIF.

In the Spanish test, there was no empirical evidence for a different factor configuration in the six items detected with high DIF between the students from the focal group and the reference group. In the case of Math, from the 10 items, empirical evidence was obtained in two of them favoring the multidimensionality hypothesis, creating differences in the factor structure between the focal group (FG) and the reference group (RG).

In summary, the agreement between some of the theoretical hypotheses and the empirical psychometric evidence obtained from the two methods for detecting DIF and from the factor structure, allowed us to identify probable irrelevant attributes related to DIF for this population, with the goal of preventing DIF in future tests of Spanish as a native language and Math.

5. Conclusions

Before providing the major substantive conclusions, two premises that were assumed by this research need to be highlighted. The first one is based on the empirical evidence obtained in multiple studies, carried out particularly in the US, which have confirmed the contribution of accommodations to more adequately measure the performance of students in large scale testing (Koretz y Barton; 2003; Sireci, Scarpati y Li, 2005). That research also holds the position assumed in this study, related to the purpose of accommodations as tools to achieve higher measurement accuracy for the morphology, syntactic and literary and mathematical knowledge of students in the focal group, eliminating or minimizing the effect of their condition or disability in their performance. The second premise states that the results of any research in this topic depend on the design, sample sizes, the kind of disability and the statistical models used in the analyses. Therefore, the following findings must be

contextualized and interpreted within the theoretical and methodological frames proposed in this study.

In this frame of reference, complementing the statistical results with the judges' criteria was essential to identify possible sources of irrelevant variance in the construct measured by these tests. Also some degree of agreement was verified between the two different empirical methods of Classical Test Theory to detect DIF, being Mantel-Haenszel more sensitive.

The theoretical contribution made by this research lies on the fact that it is the first attempt to explain an extremely complex phenomenon: a set of hypotheses, supported with empirical evidence, that could explain the reasons behind DIF for students with ADHD in standardized high school achievement tests. Regarding this issue, several authors have stressed the need of telling apart the students by their diverse disabilities, since they might benefit more from certain types of accommodations and not from others, and some might achieve a greater impact than the remaining ones (Abedi et al., 2001; Koretz, 1997).

It turned out that the statistical evidence and the expert judges' explanations supported some of the theoretical hypotheses previously defined by the researchers, using the theory behind ADHD and the exploratory "wandering" phase, prior to the statistical detection of DIF with the data. This mixed methods approach proved to be helpful to identify irrelevant attributes in Math and Spanish items that could probably causes DIF.

In the specific case of ADHD, we concluded that certain irrelevant attributes, what affect the validity in the interpretation of the results, such as text length, complex and ambiguous wording and grammar, inaccuracies in the measurement of concepts, visual overload and excess of information, both in Math and Spanish, could be potential sources of DIF, contributing to put in a disadvantages those examinees. This is due to the deficiencies associated to this disability, having problems to focus their attention needed to integrally approach the item, organizing the ideas and prescribing the details. As a clarifying note, we use here the concept of irrelevant attributes in Messick's frame of reference (1989, 1995), meaning attributes that are not part of the constructs being measured and, that, therefore, should not affect students' performance in the test.

At a more general level, the study provides elements for understanding and preventing DIF in students with ADHD in the context of constructing and validating educational tests. It shows the need to triangulate the statistical-quantitative evidence with qualitative findings in order to generate empirical hypotheses that must be verified in future confirmatory and experimental studies. Koretz and Barton (2003) confirm that, despite the increase of students

with disabilities in large-scale assessment, there is limited information regarding the use of test accommodations for students with disabilities, in both elementary and secondary levels. Information is even more limited around the effects of accommodations on the scores' validity or on the students' performance and educational achievement.

One important limitation of this research is the fact that both methods used to identify DIF do not detect non-uniform DIF (when the item favors or disfavors one group depending on the ability level, for example, for students with low ability the item could be favoring group A, and, for students with high ability could favor group B), which in itself poses a great challenge in terms of generating possible explanations.

Another methodological issue pertains to the use of the Chi-Squared statistic in Mantel-Haenszel since the inequity in sample sizes between reference and focal groups could produce unstable results for the DIF indicators (Bandeira, 2003). Besides, Aguerri et al. (2010) point out that the MH procedure is greatly affected by sample size, with inflated rates of false positives for big samples.

On the other hand, the judges did not know, prior to their work, which items were identified as exhibiting DIF and which not, their first undertaking was precisely to assess this classification in the items. This assessment proved to be difficult, rendering a somewhat low degree of concordance between judges.

Finally, since these are observational data, and the items were not constructed with the purpose of testing hypotheses about possible sources of DIF with ADHD population, there is always the risk of what is called "confounding" in Statistics, i.e. that there are other characteristics that differ in the items, besides those identified in the study, and that could also be responsible for the appearance of DIF. This limitation can only be overcome using experimental designs, where the items are constructed with the purpose of testing specific hypotheses, and in which experimental and control group items only differ in those characteristics.

6. References

Abedi, Jamal, Hofstetter, Carolyn, Baker, Eva and Lord, Carol. (2001). *NAEP Math performance and test accommodations: Interactions with student language background*. Retrieved from http://www.cse.ucla.edu/products/Reports/TR536.pdf

Aguerri, María Ester, Blum, Diego, Picón, Jimena y Galibert, María Silvia. (2010). Reglas de detección del funcionamiento diferencial del ítem. Estudio del efecto del tamaño de muestra en presencia de DIF paralelo. *II Congreso Internacional de Investigación y*

Práctica Profesional en Psicología XVII Jornadas de Investigación Sexto Encuentro de Investigadores en Psicología del MERCOSUR. Facultad de Psicología - Universidad de Buenos Aires, Buenos Aires. Recuperado de http://www.aacademica.org/000-031/920

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, Estados Unidos de América: Author.

Anastasi, Anne y Urbina, Susana. (1998). *Tests psicológicos* (7ª. ed.). Juárez, México: Prentice Hall.

Angoff, William. (1993). Perspectives on Differential Item Functioning Methodology. En Paul Holland and Howard Wainer (Eds.), *Differential item functioning* (pp. 3-23). New Jersey, United States of America: Lawrence Erlbaum Associates.

Attorresi, Horacio, Galibert, María Silvia, Zanelli, Marta, Lozzia, Gabriela y Aguerri, María Ester. (2003). Error tipo I en el análisis del funcionamiento diferencial del ítem basado en la diferencia de los parámetros de dificultad. *Revista Psicológica*, *24*(2), 289-306. Retrieved from http://www.redalyc.org/articulo.oa?id=16924207

Bandeira, Wagner. (2002). *Detección del funcionamiento diferencial del ítem (DIF) en test de rendimiento. Aportaciones teóricas y metodológicas* (Tesis doctoral, Universidad Complutense de Madrid). Recuperado de http://biblioteca.ucm.es/tesis/edu/ucm-t26457.pdf

Bandeira, Wagner. (2003). Descripción de los principales métodos para detectar el funcionamiento diferencial del ítem (DIF) en el área de la evaluación educativa. *Revista de Pedagogía Bordón*, *55*(2), 177-188.

Camilli, Gregory. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedure obscure test fairness issues? In Holland, Paul and Wainer, Howard (Eds.), *Differential item functioning* (pp. 321-335). New Jersey, United States of America: Lawrence Erlbaum Associates.

Cea, M° Ángeles. (1999). *Metodología cuantitativa*: *Estrategias y técnicas de investigación social*. Madrid, España: Editorial Síntesis.

Carvajal, Jorge y Poggio, Andrew. (2006, abril). *Studying equivalence of Spanish language versions of a large scale assessment: Differential item functioning in the cognitive and affective domain.* Work present in Annual Meeting of the National Council on Measurement in Education. San Francisco, United States of America.

Donoghue, John, Holland, Paul and Thayer, Dorothy. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and Standardization measures of differential item functioning. In Paul Holland and Howard Wainer (Eds.), *Differential item functioning* (pp. 137-166). New Jersey, United States of America: Lawrence Erlbaum Associates.

Dorans, Neil and Holland, Paul. (1993). DIF detection and description: Mantel-Haenszel and standardization. In Paul Holland and Howard Wainer (Eds.), *Differential item functioning* (pp. 35–66). New Jersey, United States of America: Lawrence Erlbaum Associates.

Elosúa, Paula y López, Alicia. (1999). Funcionamiento diferencial de los ítems y sesgo en la adaptación de dos pruebas verbales. Revista *Psicológica, 20*, 23-40. Retrieved from http://www.uv.es/revispsi/articulos1.99/elosua.pdf

Gómez-Benito, Juana, Hidalgo, María Dolores y Guilera, Georgina. (2010). El sesgo de los instrumentos de medición. Tests justos. *Papeles del Psicólogo*, *31*(1) 75-84. Recuperado de http://www.cop.es/papeles

Gómez-Benito, Juana, Balluerka, Nekane, González, Andrés, Widaman, Keith F. and Padilla, José Luis. (2017). Detecting differential item functioning in behavioral indicators across parallel forms. *Psicothema*, *29*(1), 91-95. doi: 10.7334/psicothema2015.112

Hidalgo Montesinos, María Dolores, López Pina, José Antonio y Sánchez Meca, Julio. (1997). Error tipo I y potencia de las pruebas chi-cuadrado en el estudio del funcionamiento diferencial de los ítems. *Revista de Investigación educativa*, *15*(1), 149-168.

Hidalgo Montesinos, María Dolores, Galindo Garre, Francisca, Inglés Saura, Cándido José, Campoy Menéndez, Guillermo y Ortiz Soria, Beatriz. (1999). Estudio del funcionamiento diferencial de los ítems en una escala de habilidades sociales para adolescentes. Revista *Anales de psicología*, *15*(2), 331-342. Retrieved from http://www.um.es/analesps/v15/v15_2pdf/17v98_14mdhidalg.PDF

Koretz, Daniel. (1997). The assessment of students with disabilities in Kentucky. Retrieved from http://cresst.org/wp-content/uploads/TECH431.pdf

Koretz, Daniel y Barton, Karen. (2003). *Assessing students with disabilities: Issues and evidence*. Retrieved from http://research.cse.ucla.edu/reports/TR587.pdf

Longford, Nicholas, Holland, Paul y Thayer, Dorothy. (1993). Stability of the MH D-DIF Statistics Across Populations. In Paul Holland and Howard Wainer (Eds.), *Differential item functioning* (pp. 255–276). New Jersey, United States of America: Lawrence Erlbaum Associates.

Messick, Samuel. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Research*, *18*(2), 5-11.

Messick, Samuel. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and practice, 14*(4), 5-8.

Ministerio de Educación Pública (MEP). (1998). *Políticas, normativa y procedimientos para el acceso a la educación de los estudiantes con necesidades educativas especiales* (Reimpresión de la 1ª. ed.). San José, Costa Rica: Author.

Montero, Eiliana. (1993). Linguistic and cultural influences on differential item functioning for Hispanic examinees in a standardized secondary level achievement test (Unpublished doctoral thesis in Educational Research). The Florida State University, Tallahasse, Florida, USA.

Moreno Oliver, Francesc. (2001). *Análisis psicopedagógico de los alumnos de educación secundaria obligatoria con problemas de comportamiento en el contexto escolar* (Tesis Doctoral, Universitat Autònoma de Barcelona). Retrieved from http://www.tdx.cat/handle/10803/5411

Muñiz, José. (1990). *Teoría de Respuesta a los Ítemes*. Madrid, España: Ediciones Pirámide S.A.

Padilla, José Luis, González, Andrés y Pérez, Cristino. (1998). Diferencias instruccionales y funcionamiento diferencial de los ítems: Acuerdo entre el método Mantel – Haenszel y la regresión logística. Revista *Psicológica*, *19*, 201-215. Retrieved from http://www.uv.es/revispsi/articulos3.98/padilla.pdf

Penfield, Randall. (2013). *DIFAS 5.0. Differential item functioning analysis system. User's Manual.* Retrieved from http://soe.uncg.edu/wp-content/uploads/2015/12/DIFASManual_V5.pdf

Penfield, Randall and Camilli, Gregory. (2006). Differential Item Functioning and Item Bias. In S. Sinharay and C.R. Rao (Eds.). *Handbook of Statistics*. *Psychometrics* (Vol. 26; pp. 125-167). Amsterdam, Holanda: Elsevier.

Prieto, Gerardo. (2013). *Análisis del Funcionamiento Diferencial de los Ítems de una prueba de Comprensión Lectora del Español como segunda lengua*. Recuperado de http://www.alte.org/attachments/pdfs/files/conferencia_gpa_qwroz.pdf

Rubiales, Josefina, Bakker, Liliana, Russo, Daiana and González, Rocío (2016). Desempeño en funciones ejecutivas y síntomas comórbidos asociados en niños con Trastorno por déficit de atención con hiperactividad (TDAH). *Revista CES Psicología, 9*(2), 99-113, doi: http://dx.doi.org/10.21615/cesp.9.2.7

Sireci, Stephen G., Scarpati, Stanley E. y Li, Shuhong. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457-490.

Thompson, Sandra, Blount, Amanda y Thurlow, Martha. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Technical Report 34). Recuperado de https://nceo.umn.edu/docs/OnlinePubs/TechReport34.pdf

Villalobos, Ericka y Morales, Krissia. (2002). *Niños con déficit de atención: Orientación a padres y docentes.* San José, Costa Rica: Editorial Universidad Estatal a Distancia.

Wainer, Howard. (1993). Model-Based Standardized Measurement of an Item's Differential Impact. In Paul Holland and Howard Wainer (Eds.), *Differential item functioning* (pp. 123–135). New Jersey, United States of America: Lawrence Erlbaum Associates.

Zieky, Michael. (1993). Practical questions in the use of DIF statistics in test development. In Paul Holland and Howard Wainer (Eds.), *Differential item functioning* (pp. 337-347). New Jersey, United States of America: Lawrence Erlbaum Associates.