Running head: D**evelopmental trajectories of relational concepts**

On the developmental trajectories of relational concepts among children and adolescents with intellectual disability of undifferentiated etiology

Bruno Facon
Université de Lille, SCA-LAB, France

David Magis
University of Liège, Belgium

Yannick Courbois
Université de Lille, PSITEC, France

Acknowledgement
We are grateful to John M. Belmont for his helpful comments on this manuscript. We also thank the students and psychologists who helped with data collection. We extend our deepest gratitude to the special education facilities and schools that permitted us to conduct this study, and to all the children and adolescents who participated. This work was supported by a grant from the French National Research Agency (Agence Nationale de la Recherche -ANR, LANG & HANDICAPS, Projet no. ANR-09-ENFT-019). None of the authors have any conflicts of interest with any company or other organization regarding the material discussed in the manuscript.

E-mail corresponding author : bruno.facon@univ-lille.fr

Abstract

The aim of this study was to examine the developmental trajectories of comprehension of relational concepts among 557 participants with intellectual disability (ID) of undifferentiated etiology (*M* age = 12.20 years, *SD* = 3.18) and 557 typically developing (TD) participants (*M* age = 4.57 years, *SD* = 0.80). Logistic regression analyses, with nonverbal cognitive level entered first in the equations, showed only negligible differences with regard to the discriminative power of each of the 72 concepts used as outcome variables, and moderate differences in difficulty for only 3 items. A moderate mixed effect (i.e., combining a group difference in difficulty and discriminative power) was observed for a fourth item. It is concluded that the developmental trajectories of relational concepts are similar for participants with or without ID. The implications and limitations of the study are discussed.

Keywords: Intellectual disability, item analysis, relational vocabulary, differential item functioning, developmental trajectory.

The study of intellectual disability (ID) has undergone a major change in the last decade with emphasis focused increasingly on the evolving nature of phenotypes and, therefore, reaffirming the importance of development in the study of people with ID. Indeed, until recently, the concept of phenotype has often had a static connotation, as if the peculiar characteristics of each person with ID were stable over time and, thus, would necessarily be observed at any and all points of development. This "static" view of psychological profiling (Karmiloff-Smith, 2011; Knowland & Thomas, 2011), encouraged by single age-point matching studies (Thomas et al., 2009) is now seriously challenged by the trajectory approach towards developmental disorders, whose two main principles are (1) phenotypes evolve in the course of development, and (2) no convincing explanation for a given phenotype can be provided without tracing the developmental course of each of its components (Annaz, Karmiloff-Smith, & Thomas, 2008; Dykens, Hodapp, & Finucane, 2000; Elsabbagh & Karmiloff-Smith, 2012; Karmiloff-Smith, 1998, 2011; Knowland & Thomas, 2011; Thomas, Purser, & van Herwegen, 2011). From a methodological standpoint, the developmental trajectory approach consists of building an algebraic function linking chronological age, developmental age or any other measure, with scores obtained on standardized tests, experimental tasks or neurophysiological variables. The slopesand intercepts characterizing participants of the target and control groups are then statistically compared to determine whether the developmental trajectories differ importantly (Thomas, 2016; Thomas et al., 2009).

Most studies conducted within the developmental trajectories framework can be described as “molar” in the sense that their dependent (outcome) variables are usually global scores derived from psychometric tests. Yet, even the most specific tests, that is, those designed to measure narrow dimensions of psychological development, will always have a *composite* nature. For example, despite their seemingly homogeneous content, receptive vocabulary tests such as the Peabody Picture Vocabulary Test (PPVT, Dunn & Dunn, 1997, 2007) evaluate a variety of word types. One can distinguish, for example, words belonging to different lexical or part-of-speech categories such as nouns, verbs, and adjectives (e.g., *padlock*, *share*, *furious*); root and inflected words (e.g., *mason*, *fragment* vs. *cluttered*, *lubricated*); concrete versus verbally defined words (e.g., *shrub*, *river* vs. *complaint*, *nutritive*); and basic, superordinate, and subordinate nouns (e.g., *car* vs. *vehicle* vs. *ambulance*). Similarly, the Test for Reception of Grammar (Bishop, 2003) evaluates the comprehension of various kinds of linguistic constructions (e.g., two or three-element combinations, negative sentences, reversible active sentences, singular/plural noun inflections, reversible passive sentences, embedded sentences). Given the composite nature of tests, several latent factors inevitably contribute to the variance of the overall test score. Thus, as emphasized by many psychometricians, most achievement and aptitude tests are structurally multidimensional, with one or two dominant target dimensions acting conjointly with other dimensions (Ackerman, Gierl, & Walker, 2003; Furlow, Ross, & Gagné, 2009; Reckase, Ackerman, & Carslon, 1988).

When global scores on apparently unidimensional tests are used as dependent variables, the precision of analyses is necessarily diminished. Indeed, test scores consist, at least for binary items and before all transformations that might eventually be applied to raw scores, of a sum corresponding to the total number of items correctly answered. Thus, two participants or two groups of participants can obtain the same total raw score by passing exactly the same items. In that case, the quantitative equivalence (same total raw score) goes together with a qualitative equivalence (same item response profile). Yet things do not always coincide so neatly and can significantly blur the analysis, as for example if the participants obtain the same total raw score but with totally different response profiles. In a study of mathematical skills using a standardized test battery, O'Hearn and Landau (2007) showed that the *mean* difference was not statistically significant between a group of typically developing children (TD) and a group of participants with Williams syndrome (WS) who were individually matched for mental-age on a nonverbal intelligence test. However, O'Hearn and Landau's follow-up analyses showed significant differences in favor of TD children for some items, and significant differences favoring those with WS for other items. Without the post-hoc item analysis, the authors would have missed the phenomenon. In fact, the composite nature of psychometric tests tends, *de facto*, to decrease the variance explained by the independent variable, unless the strength of the relationship is of the same order between the independent variable and each of the latent dimensions of the test used as the dependent variable.

An interesting way to overcome this problem could be to move from a molar to a molecular level of analysis by performing statistical analyses on item responses rather than on the whole test score. For such an approach, the many analytical tools developed in the item-analysis framework could be of great help. These tools are designed to examine whether items from psychometric tests present a *differential functioning* (DIF) related, among other things, to examinees' gender, ethnic origin, socioeconomic status or linguistic background. They aim to improve the fairness of tests by ensuring that each item evaluates the construct(s) targeted by the test and not specific traits related to membership in a particular group (Camilli & Shepard, 1994; Holland & Wainer, 1993; Osterlind & Everson, 2009).

The item-analysis techniques could inform fine-grained developmental trajectory analyses at the item level, for example by comparing each individual item's characteristic curves for two (or more) groups (e.g., TD versus participants with ID) who were previously matched on the relevant developmental trait. An item's characteristic curve is the function linking the total score on a test (x-axis) to the probability of passing one of its items (y-axis). An item can be considered as functioning similarly for two groups of examinees if their characteristic curves for that item are closely similar (Figure 1, panel A). In this case, the item's discriminative power is the same (the two groups' curves have the same slope) as is its difficulty level (the two curves have the same location along the x-axis). If the characteristic curves do not closely overlap, the item is said to present a DIF, which can stem from a difference of difficulty (the probability of a correct response for one group is significantly greater than that of the other group at any point on the x-axis, Figure 1, panel B) and/or of discriminative power (difference of slope for the two curves, Figure 1, panel C). The DIF is said to be “uniform” in the first case and “nonuniform” in the second. Of course, differences in difficulty (uniform DIF) and discriminative power (nonuniform DIF) can appear simultaneously.

In the developmental trajectories framework, a uniform DIF would mean that the rate of acquisition of the trait evaluated by the item is the same for both groups (characteristic curves with similar slopes) but that one of the two groups shows a learning delay with respect to the other (different location of characteristic curves along the x-axis). On the other hand, a nonuniform DIF would mean that the rate of development of the trait is greater for one group than the other (characteristic curves with different slopes) and, consequently, that the developmental trajectories diverge.

In the present work, we used the item-analysis approach to compare the acquisition rate of relational concepts of participants with or without ID. Relational vocabulary must be distinguished from general vocabulary (Facon, Magis, & Courbois, 2012; Fazio, Johnston, & Brandl, 1993; Mervis & John, 2008; Miolo, Chapman, & Sindberg, 2005). General vocabulary, called “concrete” vocabulary by some researchers (e.g., Mervis & John, 2008), comprises mainly nouns, verbs and adjectives referring to objects, actions, events, states or processes. The most well-known test of this kind of vocabulary is the PPVT (Dunn & Dunn, 1997, 2007). Conversely, relational vocabulary consists exclusively of abstract words indicating spatial, temporal, dimensional, quantitative or class relationships between objects, persons or events, such as "behind", "third", "inside", "larger", "before, "in front of", or "never". Sometimes called "basic concepts", these terms are more difficult to comprehend and produce for the child because the terms have less stable and less tangible relationships with their referents (Boehm, 2000).

Very few studies have been conducted on the development of the relational lexicon among persons with ID. However, a fairly safe conclusion is that the sequence of acquisition of these words is similar for participants with or without ID (e.g., Facon et al., 2012). A recent study also showed that the developmental trajectories of relational concepts of participants with Down syndrome, participants with ID of undifferentiated etiologies and TD children matched on nonverbal intelligence level were wholly the same (Facon, Courbois, & Magis, 2016). In that study, however, the outcome variable was a composite measure including items evaluating concepts of space, time, number or quantity. Thus, the study was limited by the above-mentioned methodological shortcoming. In particular, one cannot know from its findings whether the trajectory of acquisition of each concept taken separately is similar for participants with or without ID. We here address this issue by using multiple logistic regression analysis (Swaminathan & Rogers, 1990). Specifically, we examined the mastery of 72 relational concepts from the Boehm Test of Basic Concepts (BOEHM, Boehm, 2009a, 2009b) among participants with or without ID by successively entering, in each of the 72 regression equations, their score on a nonverbal intelligence test (Raven's Colored Progressive Matrices, [RAVEN] Raven, Court, & Raven, 1998), their diagnostic status (with or without ID) and the interaction term (nonverbal developmental level × diagnostic status). The degree of overlap of logistic curves will indicate whether the developmental trajectory of each concept is similar or different across the two groups.

Method

Participants

There were two groups of participants tested as part of a larger study on language development of persons with ID supported by the French National Research Agency and for which the Ethics Committee of the Cognitive and Affective Sciences Laboratory (SCALab, University of Lille) had granted ethical approval. The first group included 557 TD participants (*M* age = 4.57 years, *SD* = 0.80) recruited in 47 regular kindergartens or elementary schools, none of whom had ever been referred for a psychological assessment at school. The second group included 557 participants with ID (*M* age = 12.20 years, *SD* = 3.18) enrolled in 51 special education schools for youngsters with mild to severe ID This group was composed of persons with ID of unknown origin or persons with ID of a wide variety of known causes (i.e., genetic syndromes, fetal alcohol syndrome, pre- or perinatal brain injuries, infectious diseases, etc.). All participants included in the study came from a French-speaking families.

TD participants were exactly matched with participants with ID using their RAVEN raw scores. The aim of this matching was to make the distribution of nonverbal cognitive levels exactly the same regardless of diagnostic status. Thus, if differences in trajectories are observed between TD participants and those with ID for the mastering of relational concepts, the shape of the distributions of RAVEN scores could not be invoked as a potentially confounding factor (see, Facon, Magis, & Belmont, 2011).

Descriptive statistics for chronological age, gender, the RAVEN and the BOEHM are given in Table 1. Each group's distribution of chronological ages is shown in Figure 2. Because of the matching procedure, the difference between the two groups' mean RAVEN scores was nonsignificant ($t_{2\text{-}tailed} = 0.000$, df = 1112, $p = 1.00$), as was the Levene test for homogeneity of variance ($F_{(1,1112)} = 0.000$, $p = 1.00$). The means were also very similar for the total score on the BOEHM ($t_{2\text{-}tailed} = -\ 0.560$, df = 1112, $p = .576$) and, although the dispersion of scores on this test was wider for the participants with ID, the Levene test for homogeneity of variance was not quite significant at $\alpha = 0.05$ ($F_{(1,1112)} = 3.150$, $p = .076$).

To check the quality of the matching on nonverbal cognitive level, the percentage of correct responses on each item of the RAVEN was computed for the two groups. The correlation between the two series of 36 percentages was .98 ($p < .000001$). Participants with and without ID are therefore matched on their whole test score *as well as each* item score. This almost perfect correlation means that the underlying cognitive processes are presumably the same for the two groups (see, Facon & Nuchadee, 2010). Factor analysis of item scores identified two factors of very similar nature for the two groups: The correlation of the 36 saturations–TD vs ID–was .89 ($p < .000001$) for the first factor and .77 ($p < .000001$) for the second.

The correlations between chronological age, the RAVEN, the BOEHM, and gender of participants of each group appear in Table 2. For TD participants, the correlations between chronological age, the RAVEN and the BOEHM were moderate to high, which was not surprising from a developmental perspective. There was also a strong correlation between RAVEN and BOEHM scores for participants with ID, a fact that could also be anticipated given the link between language and cognition. However, even if they were significant due to the large sample size, the correlations between chronological age and scores on the RAVEN and the BOEHM of participants with ID are negligible (.097 and .104, respectively). This results from the cross-sectional character of the study design and the matching procedure used to form the groups. In a longitudinal study, chronological age of participants with ID would necessarily have been correlated with their nonverbal cognitive level. However, the very low correlation found here is crucial for the present study. Indeed, if between-groups differences of characteristic curves (see below) of BOEHM test items were observed, they could not be interpreted as a chronological age-related effect. Likewise, the negligible relationship between chronological age and the

RAVEN score means that the severity of intellectual disability of participants with ID is not correlated with the RAVEN score.

Finally, because correlations between gender and all other variables approach zero, participants' gender could not be invoked as a causal factor when interpreting the results.

Instruments

The *Test des Concepts de Base* (BOEHM, Boehm, 2009a, 2009b — the French version of the Boehm Test of Basic Concepts) and Raven's Colored Progressive Matrices (RAVEN, Raven, Court, & Raven, 1998) were individually administered with no time limits by master's students in developmental psychology or contract psychologists trained in psychometrics. Testing sessions were conducted in quiet rooms situated near participants' classrooms.

For each item of the BOEHM, 4 to 6 options are displayed on a page and the participant must select the one corresponding to a concept spoken by the examiner. This test evaluates only abstract words indicating spatial, temporal, dimensional, quantitative or class relationship. The BOEHM is available in two French-language versions, one for *preschool* (Boehm, 2009a) and the other for *kindergarten to 2nd grade* (Boehm, 2009b). The *Preschool* version, intended for children ages 3 to 5 years 11 mo, comprises 76 items designed to evaluate 38 concepts (2 items per concept). The *Kindergarten to 2nd grade* version applies to children ages 5 to 8 years. It comprises 50 items each evaluating a particular concept. For the study, the two versions of

Boehm's test were combined into a single test that was individually administered to each participant. This was done to avoid the inevitable floor and ceiling effects that would occur using only one or the other test. To reduce test duration, one item from each conceptual pair of the *Kindergarten* version was deleted, as was one from each pair of items that were duplicated across the two versions. The final test comprised 72 items administered according to the order recommended in the original test manuals. This modified version of the test was used in a recent study conducted with participants with and without ID. In that study, reliability coefficients approached .90 and the rank order difficulty of items was very similar across the two types of participants (Facon et al., 2012). The α-Cronbach coefficients computed on the present study's data also indicate a very high reliability (Table 1).

The RAVEN, a well-known nonverbal intelligence test for children, was administered to all participants to obtain an estimate of their cognitive level. Each of the 36 items is presented as a colored pattern with a missing portion and 6 options for filling in the missing element. This test was chosen because of the simplicity and speed of its administration and scoring, its reliability, and the great similarity of item response profiles to which it gives rise for participants with and without ID (Facon & Nuchadee, 2010; Facon, Magis, Nuchadee, & De Boeck, 2011). Moreover, the RAVEN is used extensively to assess the fluid-like component of intelligence of typical and clinical populations of children (Cotton et al., 2005).

Statistical analyses

A logistic regression analysis was conducted for each of the 72 items of the BOEHM to estimate the contributions of the RAVEN, the participant's diagnostic status and the RAVEN × status interaction. The RAVEN was entered first in the equations. In this way, the nonverbal cognitive level cannot be invoked as a causal variable if a difference in characteristic curves is observed between the two groups. The status variable, coded 1 or 0 for participants with or without ID, respectively, was then entered followed by the interaction term. A main effect of diagnostic status would indicate a systematic difference in response probability across groups corresponding to a uniform DIF. In this case, the probability of a correct response for one group will be greater than that of the other group at all points on the x-axis (see Figure 1, panel B). On the other hand, a significant interaction would indicate a between-groups difference in slopes of logistic curves and, thus, a difference in the acquisition rate of the concept. In the item-analysis framework, an interaction effect corresponds to a difference of item discriminative power, which is a nonuniform DIF (see Figure 1, panel C).

For each item, the increase of the squared multiple correlation coefficient (ΔR^2) upon the introduction of clinical status and the RAVEN × status interaction in the regression equation was computed and statistically tested to obtain an estimate of the effect size of each of these two variables for each of the 72 items.

By DIF effect, one usually means the difference in the probabilities of answering an item correctly by two or more groups of participants when the ability level is held constant. In many DIF studies, the ability level is an *internal criterion* (i.e., the total score on the test from which the item is derived) that is used to control for the ability level of participants (Osterlind & Everson, 2009). In the present study, choosing this option would have led to using the BOEHM score instead of the RAVEN in the regression equations. Another option is to use an *external criterion*, which is an ability measure of a different sort from that of the items under study. In the present work, we opted for using an external criterion (viz., the RAVEN score) for two main reasons. The first is that nonverbal cognitive tests are frequently used to control for developmental level in studies on language acquisition of children with ID. The second was to avoid the criticism of circularity that can be leveled at studies that use an internal ability criterion (see, Camilli & Shepard, 1994).

R (R Development Core Team, 2017) was used for fitting the logistic models and related statistical tests and computations. Given the number of comparisons, the type I error rate was controlled using the False Discovery Rate (FDR) described by Benjamini and Hochberg (1995) because, compared to other adjustment methods for multiple comparisons (e.g., the Bonferroni correction), it allows control of the type I error rate with a reduced impact on statistical power. In other words, the FDR solution is less conservative than Bonferroni's and will therefore limit the number of false negatives. For more details on the mathematical foundations of the approach, see Benjamini and Hochberg (1995) and, for a very accessible presentation, McDonald (2014).

Given the introduction of the RAVEN score in the regression equations to control for the effect of nonverbal ability level on the probability of passing each BOEHM item, the prior matching of groups may seem unnecessary. However, we judged it methodologically relevant because the one disadvantage of the logistic regression DIF detection approach is that between-group differences in means or dispersions of ability levels increase the type I error rate (Pei & Li, 2010; Sireci & Rios, 2013). This is because the data density is not the same for the two groups along the ability continuum, thus making it problematic to estimate the parameters of the regression equation.

According to Zumbo (1999), a sample size of 200 participants per group is adequate for DIF studies using the logistic regression method. However, simulation studies show that 500 to 600 participants or more per group considerably increase the statistical power of the analyses (e.g., Finch & French, 2007; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Whitmore & Schumacher, 1999). From this standpoint, the present study can be considered as sufficiently powered to detect DIF items. Of course, high statistical power increases the risk of flagging items with statistically significant p-values but practically trivial effect sizes. To avoid this problem, we used the guidelines proposed by Jodoin and Gierl (2001) in which the effect size can be considered as *negligible* if $\Delta R^2 < .035$, *moderate* if $.035 \leq \Delta R^2 < .07$, or *large* if $\Delta R^2 \geq .07$. Indeed, these guidelines are generally judged as more adequate than others—such as those proposed by Zumbo and Thomas (1997) — in DIF studies conducted within the logistic regression approach (French & Maller, 2007; Gómez-Benito, Hidalgo, & Padilla, 2009).

Results

The group effect, which indicates a group difference in item difficulty, was significant for 29 of the 72 items of the BOEHM (40%) using the FDR correction for multiple comparisons. About half of these 29 items (52%) were more difficult for the participants with ID, a result which could be anticipated given the nonsignificant difference of mean raw scores of the two groups on this test (Table 1). Without correction of p-values for multiple comparisons, 34 differences would have been significant, but only 17 with Bonferroni's correction. However, beyond the number of p-values, and given the large sample size, it is mainly the effect sizes that matters. From this standpoint, the results are much less conclusive. Indeed, the proportion of variance explained by diagnostic status is almost always negligible. For the 72 tested items, no large effects ($\Delta R^2 \geq .07$) and only 3 moderate effects ($.035 \leq \Delta R^2 < .07$) were observed (Figure 3). The latter were :

- Item 11 ("*Montre-moi les jouets qui sont* ***à l'extérieur*** *de la boîte*" = "*Show me the toys that are* ***outside*** *the box*" [TD < ID]) ;

- Item 27 ("*Montre-moi la fille qui est* ***avant*** *le garçon dans la file" = "Show me the girl who is* ***in front of*** *the boy in line*" [TD > ID]) ;

- Item 64 ("*Regarde les enfants et la corde. Montre-moi l'enfant qui saute* ***par-dessus*** *la corde*" = "*Look at the children and the rope. Show me the child who is jumping* ***over*** *the rope*" [TD > ID]).

To complete the information on the difficulty of items and to allow for comparisons with results of future studies, the percentages of correct responses on each item along with their rank order of difficulty are presented for each group separately in the Appendix.

As stated previously, a significant interaction term indicates a nonuniform DIF, that is, a between-groups difference in the item's discriminative power. The RAVEN × status interaction was significant for only one item. The discriminative power of that item was slightly better for the TD participants. However, the variance explained by the 72 interaction terms was *always* negligible. In fact, *all* the ΔR^2 were less than .0146, which is far below the threshold set by Jodoin and Gierl (2001) to distinguish between negligible and moderate effect sizes (Figure 4). In other words, none of the BOEHM items exhibited a compelling nonuniform DIF. Thus, one may conclude that the items' discriminative power is not affected by the participant's clinical status.

Finally, a significant moderate mixed DIF effect (i.e., combining negligible group differences in difficulty and in discriminative power) was observed for item 39, whose effect size was slightly above the Jodoin and Gierl threshold (ΔR^2 = .036).

- "*Regarde les chiens qui jouent*. *Montre-moi le chien qui passe* ***à travers*** *le cerceau" =* "*Look at the dogs who are playing*. *Show me the dog that is going* ***through*** *the hoop*" ([difficulty: TD>ID; discriminative power: TD < ID]).

The characteristic curves of the 4 items flagged as showing DIF are presented in Figure 5. Beyond the moderate size of these 4 effects, it is important to emphasize their very low number. In fact, almost 95 % of BOEHM items have comparable degree of difficulty and discriminative power across the two groups; and many of them have characteristic curves that are practically indistinguishable across groups (see examples in Figure 6).

To extend the scope of the analysis and to show that the present results were not the consequence of the use of an external criterion (i.e., a nonverbal intelligence test) to control the developmental level of participants, the statistical analysis just described was replicated using an internal criterion (viz., the total score on the BOEHM) as the matching variable. As the average scores of participants with and without ID were almost comparable on the BOEHM ($t_{2\text{-tailed}}$ = -0.530, df = 1144, p =.596), logistic regression analyses were conducted without changing the composition of the groups. In these analyses, the BOEHM test score, the clinical status of participants and the interaction term (BOEHM × status) were successively entered into the equations.

The results of this second statistical analysis almost completely corroborated those of the first. Indeed, only 5 items were flagged as showing a moderate DIF, three of which had already been detected during the first analysis (items 11, 27, and 64). The fourth (item 54) showed a uniform DIF and the fifth (item 41) a mixed DIF effect (i.e., combining negligible group differences in difficulty and discriminative power). These two items were:

- Item 41 ("*Regarde les t-shirts. Montre-moi le t-shirt qui est de taille moyenne" = "Look at the t-shirts. Show me the t-shirt that's medium size*." [difficulty: TD>ID; discriminative power: TD > ID]) ;
- Item 54 ("*Regarde le lapin, le chat, le cochon et le chien. Montre-moi l'animal qui est près du lapin*." = "*Look at the rabbit, the cat, the pig and the dog. Show me the animal that's near the rabbit*." [TD > ID]).

To show the consistency of results from the two analyses, the proportions of variance explained both by the clinical status of participants and the interaction term from the first analysis (with the RAVEN score used as the measure of developmental level) were compared with those obtained in the second analysis (with the BOEHM score used as the measure of developmental level). Results showed that the two analyses were highly consistent. Indeed, the paired points from the two ΔR^2 sets formed a narrow ellipse (Figure 7) and their correlation was close to unity (r = 0.95, $p. < .00001$). Finally, as further proof of the consistency of the two analyses, items with the same characteristic curves for participants with or without ID in the first analysis (see the four examples provided in Figure 6) also have almost totally superimposed curves in the second (Figure 8).

Discussion

One of the methodological difficulties encountered in the study of developmental trajectories arises from the composite nature of measures often used as outcome variables. In the present study, we attempted to move from a molar to a molecular level of analysis by examining, within the methodological framework of item analysis, the developmental trajectories of each concept included in the BOEHM test. Results of logistic regression analyses were clear-cut. Only 4 items were flagged as DIF when the two groups were matched on the RAVEN score, and only 5 when the BOEHM test score was used as the matching variable. Therefore, it can be concluded that the developmental trajectories of concepts evaluated by the BOEHM are similar for the study's participants with or without ID. This conclusion is valid for participants with or without ID who are matched on developmental level. If the matching had been done on chronological age, all trajectories would have differed. Such matching was not attempted in this study because of the large between-groups chronological age differences and, consequently, the almost total absence of overlap of the two age distributions.

Despite a detailed inspection of the three items flagged as DIF in the two analyses (item 11 [*outside,* TD < ID], 27 [*in front of,* TD > ID], and 64 [*over,* TD > ID]), we have not identified factors that might explain this result. We hypothesize that the divergent developmental trajectories observed for these 3 items are the consequences of different educational experiences, but we are unable to say which aspects of these experiences might be in play.

The absence of differences between the two groups of participants cannot be interpreted as the consequence of a statistical inability to separate the “signal” from the “noise”, for example, because of a lack of reliability of BOEHM’s test items. It is true that a participant's score on an item is necessarily less reliable than that obtained on a test containing 30, 40 or 50 items. However, if BOEHM's test items were not reliable, the reliability coefficients computed for the overall score would themselves be very low due to the strong relationship between the measurement error of individual items and the total-score measurement error. Yet, Cronbach alpha coefficients are particularly satisfactory for the two groups and, therefore, allow us to conclude that items from this test are themselves reliable. Moreover, the common denominator of the various statistical approaches for detecting DIF between groups (e.g., Camilli & Shepard, 1994; Magis, Béland, Tuerlinckx, & De Boeck, 2010; Osterlind, 1983; Osterlind & Everson, 2009; Penfield & Camilli, 2007; Sireci, Patsula, & Hambleton, 2005) is the use of large samples of participants. The larger the sample, the lower the measurement error at the item level and the higher the reliability of the measure. Thus, in accordance with the law of large numbers, the *empirical* mean score of a group of participants on a given test item converges towards its *true* mean score as the sample size increases. This is why, in the present study, we constituted two large samples, which reduced the measurement error on each item score. Finally, if there were a statistical inability to separate the "signal" from the "noise", the probability of success on BOEHM's items would not increase with the nonverbal cognitive level of participants. It would also be difficult to explain why many of the item's logistic curves are almost indistinguishable across groups regardless of which test is used to control for developmental level (see Figures 6 and 8).

Another potential problem concerns the young age of TD participants for whom testing could be an unusual and potentially destabilizing situation. This lack of testing experience might not have allowed them to accommodate with the requirements of tests such as Raven's matrices. As a result, their nonverbal cognitive level would not have been properly assessed, which could explain the negligible between-groups differences of developmental trajectories. That seems unlikely, though, because the tests' reliability coefficients are rather satisfactory for each group. In addition, the correlation between the proportion of correct responses of participants with and without ID for the 36 items of the RAVEN is close to unity (.98) and the factorial structure of this test is very comparable for both groups. Finally, there are only very few between-groups differences of difficulty and discriminative power for the BOEHM test items *regardless* of the measure used to match the groups.

One implication of these results is that tests of relational vocabulary are appropriate for assessing children with ID. Controlling for cognitive level, the difficulty and discriminative power parameters of the BOEHM test items found in the ID group's performance are almost identical to those observed among the TD children. Thus, these items do not present a differential functioning and so cannot be seen as disadvantaging one group or the other. This conclusion may well apply to other tests of relational concepts such as the Bracken Basic Concept Scale (Bracken, 2006) or the Test of Relational Concepts (Edmonston & Litchfield-Thane, 1988). It is possible, however, that the conclusion will not hold for individuals from some particular genetic syndromes (see below).

Another implication concerns pedagogical strategies to promote the acquisition of relational concepts. Given the similarity of developmental trajectories observed across the two groups, one may consider that concept-learning programs devised for typical or at-risk children (Bereiter & Engelmann, 1966; Boehm, 1976; Bracken, 1986; Hansen, 2009) can be used without major adaptations with children with ID. This does not mean, however, that adaptations are not to be considered, particularly for children from specific etiological groups (see below).

A third implication is that processes underlying the acquisition of relational vocabulary are robust in that they do not appear to be affected by ID. Indeed, apart from the delay, here clearly highlighted —the chronological age difference between the two groups is about 8 years, but the average scores on the BOEHM test are nearly the same— the developmental pathway of almost all investigated concepts seems non-specific. In this respect, the present work confirms the results of other research showing that the vocabulary development of children with ID is far more a matter of delay than difference (e.g., Berglund, Eriksson, & Johansson, 2001; Facon et al., 2016; Facon et al., 2012; Facon, Nuchadee, & Bollengier, 2012; Grela, 2002; Hart, 1996; Loveall, Channell, Philipps, Abbeduto, & Conners, 2016; Philipps, Loveall, Channell, & Conners, 2014; Polišenská & Kapalková, 2014; Polišenská, Kapalková, & Novotková, 2018).

This work requires further development targeting other components of language such as general (or concrete) lexicon, syntax, phonology or pragmatics. This would extend the current results which cover only a limited aspect of language development. Indeed, it is possible that differences in difficulty and/or discriminating power of items may be observed for tests other than the BOEHM. In this respect, it has been shown that when participants with and without ID are matched on their overall developmental age with a composite intelligence scale, different profiles of abilities can be observed. TD participants are generally better on tasks involving verbal reasoning, speed of processing and abstraction. By contrast, participants with ID surpass them on target tasks involving chronological age-related learning products, that is, to the educational experience accumulated over the years (e. g., Baughman, Thomas, Anderson, & Reid, 2016; Blount, 1970; Cruickshank & Qualtere, 1950; Eaton & Burdz, 1984; Fazio et al., 1993; Hore & Tryon, 1989; Martinson & Strauss, 1941; Meyers, Dingman, Attwell, & Orpet, 1961; Santucci & Hélal, 1969; Spitz, 1982). Beyond composite intelligence test profiles, this age-related experience effect has also been shown for scores on general receptive vocabulary tests, which often exceed nonverbal cognitive measures for individuals with ID, particularly in late childhood and adolescence (Chapman, 2006; Facon, Bollengier, & Grubar, 1994; Facon & Facon-Bollengier, 1997, 1999; Facon, Facon-Bollengier, & Grubar, 2002; Facon, Grubar, & Gardez, 1998; Miolo et al., 2005). In an item analysis study, we might therefore expect, for tests of specific components of language development, to discover a significant number of items showing differential functioning of moderate and even large effect size for groups of participants with or without IDs matched on nonverbal cognitive level. However, this remains to be empirically demonstrated.

Concerning generalization, it would be appropriate to take account of the etiology of participants with ID, which was not done in the current study. Thus, it cannot be concluded that the present results are universally valid for well-defined syndromes such as Down, Fragile X or Williams (WS). Indeed, a growing number of studies of cognitive, behavioral, and emotional phenotypical features of people with ID have shown that etiology has specific effects on the structure and functioning of the brain and, thereby, on the psychological phenotype (e.g., Jonas, Montojo, & Bearden, 2013; Lightbody & Reiss, 2009). Therefore, ID should not be studied without grouping participants by etiology (Fidler, Daunhauer, Gerlach-McDonald, & Schworer, 2016). Initially focused on a few known etiological groups (e.g., Down, Williams or Fragile X syndromes), the syndromic approach has been extended to an increasing number of syndromes such as 22q11.2 deletion (Biswas & Furniss, 2016), 7q11.23 locus duplication (Somerville et al., 2005), fetal alcohol (Kingdon, Cardoso, & McGrath, 2016), Prader-Willi (Griggs, Sinnayah, & Mathai, 2015), Angelman (Mertz et al., 2014), Wolf-Hirschhorn (Fish et al., 2012) or Smith-Magenis (Alaimo et al., 2015). However, this approach is precluded for many rare syndromes by the paucity of available participants. What is gained by homogenizing etiology is lost because of reduced statistical power and analytic precision. Given the relationship between sample size and statistical power (e.g., Krzywinski & Altman, 2013), there is an increase in type II errors with small samples, meaning an increase in falsely rejected alternative hypotheses. Small samples thus raise doubts about non-significant results, which may be attributed to a lack of effect or equally to a lack of power. Had the current study involved only, say, 30 participants per group with similar results (i.e., almost no significant between-group differences), the reader would justifiably attribute the absence of effects to a lack of power. Therefore, it is always necessary to privilege statistical power even sometimes at the likely expense of etiological purity. This power problem in ID research arises from the fact that many ID-related genetic syndromes occur in the range of 1/10,000 to 1/50,000 of the population (McKusick-Nathans Institute of Genetic Medicine, 2019). To achieve adequate numbers of matched participants for an item analysis study (say 450 per group, one of which represents a live-birth rate of 1/10,000) one would run multi-site, even multicountry collaborative studies on existing test result databases. With 500 million inhabitants in the European Union, and a yearly birthrate of ~12 per 1000, there would be ~600 babies born per year with the target etiology. Only 60 / year would be needed to populate an 8-year study involving 450 total participants in the target etiology group. Thus, insofar as the same tests are often used at different sites by independent research teams within a given country, the pooling of item responses of participants with specific etiologies could yield sample sizes sufficient for fine-grained item analyses even for relatively rare syndromes. In this respect, initiatives such as the "Psychological Science Accelerator" might prove to be promising (see, Moshontz et al., 2018).

The need for further item-analysis studies with better control of etiology is well illustrated by research on spatial vocabulary of children with WS. Several studies have shown that visual and spatial difficulties of people with WS (e.g., Farran & Jarrold, 2003; Mervis, Morris, Bertrand, & Robinson; 1999) result in specific difficulties in the mastering of spatial concepts (Bellugi, Lichtenberger, Jones, & Lai, 2000; Phillips, Jarrold, Baddeley, Grant, & Karmiloff□Smith, 2004). However, other studies have shown that beyond spatial concepts, *all* relational concepts are affected among participants with WS (e.g., Mervis & John, 2008). By combining item performances of participants with WS on the same test of relational concepts (e.g., The Boehm Test of Basic Concepts [Boehm, 2000], the Bracken Basic Concept Scale [Bracken, 2006] or the Test of Relational Concepts [Edmonston & Litchfield-Thane, 1988]) gathered by different research teams working on WS, sample sizes would be sufficient to yield adequate statistical power and thus to determine whether or not the developmental trajectories of relational concepts of participants with WS are comparable to those of TD children. This type of research could be replicated with other etiological groups and with participants with autism spectrum disorders. The present study shows that even without taking account of the etiology of ID, it appears that intellectual deficiency does not lead to group-specific developmental trajectories for relational vocabulary. This is a first step towards more advanced research with a greater focus on etiology of ID.

A further limitation of the study is the lack of data on parental education and socioeconomic status (SES). As these are related to language development among TD children (e.g., Fernald, Marchman, & Weisleder, 2013; Hart & Risley, 1995; Hoff, 2013) and those with ID (Price, Roberts, Vandergrift, & Martin, 2007; Warren, Brady, Sterling, Fleming, &Marquis, 2010), these variables might stand as informative covariates in future studies of developmental trajectories of language components of persons with ID. Fortunately for the present study, SES and parental level of education were indirectly controlled by matching participants on the level of nonverbal cognitive development and then on the level of relational vocabulary.

The cognitive processes involved in item responses also remain to be investigated. Indeed, the similarity of trajectories of concepts acquisition of participants with or without ID does not necessarily mean that the processes involved are the same. What appears unaltered, intact or similar in spite of ID could possibly be something different resulting from a reorganization of the whole cognitive / linguistic system (Karmiloff-Smith, Brown, Grice, & Paterson, 2003; Richardson & Thomas, 2009). This possibility will be sorted out only by targeted laboratory studies.

In conclusion, the present findings do not indicate different developmental trajectories of relational concepts among participants with or without ID. However, although they seem solid in view of the methodology used and the large sample sizes, the scope of the study remains limited to one specific aspect of language development. Further studies are needed to flesh out our knowledge of other components of language (lexical, syntactic or phonologic), whether in reception or in production. Moreover, although the approach used in the present work allows fine-grained analyses, solutions still need to be found for conducting comparable research with specific etiological groups.

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*, 37-53. doi:10.1111/j.1745-3992.2003.tb00136.x

Alaimo, J. T., Barton, L. V., Mullegama, S. V., Wills, R. D., Foster, R. H., & Elsea, S. H. (2015). Individuals with Smith-Magenis syndrome display profound neurodevelopmental behavioral deficiencies and exhibit food-related behaviors equivalent to Prader-Willi syndrome. *Research in Developmental Disabilities*, *47*, 27-38. doi:10.1016/j.ridd.2015.08.011

Annaz, D., Karmiloff-Smith, A., & Thomas, M. S. C. (2008). The importance of tracing developmental trajectories for clinical child neuropsychology. In J. Reed & J. Warner-Rogers (Eds.), *Child neuropsychology: Concepts, theory and practice* (pp. 7-18). Oxford, UK: Wiley-Blackwell.

Baughman, F. D., Thomas, M. S. C., Anderson, M., & Reid, C. (2016). Common mechanisms in intelligence and development: A study of ability profiles in mental age□matched primary school children. Intelligence, 56, 99□107. doi.org/10.1016/j.intell.2016.01.010

Bellugi, U., Lichtenberger, L., Jones, W., & Lai, Z. (2000). The neurocognitive profile of Williams syndrome : A complex pattern of strengths and weaknesses. *Journal of Cognitive Neuroscience*, *12*, 7-29. doi: 10.1162/089892900561959

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B*, *57*, 289-300.

Bereiter, C., & Engelmann, S. (1966). *Teaching disadvantaged children in the preschool*. Englewood Cliffs, NJ: Prentice-Hall.

Berglund, E., Eriksson, M., & Johansson, I. (2001). Parental reports of spoken language skills in children with Down syndrome. *Journal of Speech, Language, and Hearing Research*, *44*, 179-191. doi:10.1044/1092-4388(2001/016

Bishop, D. V. M. (2003). *The Test for Reception of Grammar, Version 2 (TROG-2)*. London, UK: Psychological Corporation.

Biswas, A. B., & Furniss, F. (2016). Cognitive phenotype and psychiatric disorder in 22q112 deletion syndrome: A review. *Research in Developmental Disabilities*, *53-54*, 242-257. doi:10.1016/j.ridd.2016.02.010

Blount, W. R. (1970). Retardates, normals, and U.S. moneys: Knowledge and preference. *American Journal of Mental Deficiency*, *74*, 548-552.

Boehm, A. E. (1976). *Boehm resource guide for basic concept teaching*. New York, NY: The Psychological Corporation.

Boehm, A. E. (2000). Assessment of basic relational concepts. In B. A. Bracken (Ed.), *The psychoeducational assessment of preschool children* (3rd ed., pp. 186-203). Needham Heights, MA: Allyn & Bacon.

Boehm, A. E. (2009a). Boehm 3 maternelle. *Test des concepts de base (troisième édition) [Boehm 3 kindergarten. Test of basic concepts (3rd ed.)]*. Paris: Les Editions du Centre de Psychologie Appliquée.

Boehm, A. E. (2009b). *Boehm 3. Test des concepts de base (troisième édition) [Boehm 3. Test of basic concepts (3rd ed.)]*. Paris: Les Editions du Centre de Psychologie Appliquée.

Bracken, B. A. (1986). *Bracken concept development program*. San Antonio, TX: The Psychological Corporation.

Bracken, B. A. (2006). *Bracken Basic Concept Scale–Receptive (3rd ed.)*. San Antonio, TX: Harcourt Assessments.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.

Chapman, R. S. (2006). Language learning in Down syndrome: The speech and language profile compared to adolescents with cognitive impairment of unknown origin. *Down Syndrome: Research & Practice*, *10*, 61-66. doi: 10.3104/reports.306

Cotton, S. M., Kiely, P. M., Crewther, D. P., Thomson, B., Laycock, R., & Crewther, S. G. (2005). A normative and reliability study for the Raven's Coloured Progressive Matrices for primary school aged children from Victoria, Australia. *Personality and Individual Differences*, *39*, 647-659. doi:10.1016/j.paid.2005.02.015

Cruickshank, W. M., & Qualtere, T. J. (1950). The use of intelligence tests with children of retarded mental development: II. Clinical considerations. *American Journal of Mental Deficiency*, *54*, 370-381.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test (3rd ed.)*. Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test (4th ed.)*. Bloomington, MN: NCS Pearson, Inc.

Dykens, E. M., Hodapp, R., & Finucane, B. (2000). *Genetics and mental retardation syndromes: A new look at behavior and interventions*. Baltimore, MD: Paul H Brookes Publishing Company.

Eaton, W. O., & Burdz, M. P. (1984). Gender understanding and the similar sequence hypothesis. *American Journal of Mental Deficiency*, *89*, 23-28.

Edmonston, N. K., & Litchfield Thane, N. (1988). *TRC: Test of Relational Concepts*. Austin, TX: Pro-Ed.

Elsabbagh, M., & Karmiloff-Smith, A. (2012). The contribution of developmental models toward understanding gene-to-behavior mapping: The case of Williams syndrome. In J. Burack, R. Hodapp, G. Iarocci, & E. Zigler (Eds.), *The Oxford handbook of intellectual disability and development* (pp. 30-41). New York, NY: Oxford University Press.

Facon, B., Bollengier, T., & Grubar, J. C. (1994) Déficience mentale: influence de la dissociation entre efficience et expérience. *Enfance,* 1, 71-81.

Facon, B., Courbois, Y., & Magis, D. (2016). A cross-sectional analysis of developmental trajectories of vocabulary comprehension among children and adolescents with Down syndrome or intellectual disability of undifferentiated aetiology. *Journal of Intellectual and Developmental Disability*, *41*, 140-149. doi:10.3109/13668250.2016.1160370

Facon, B., & Facon-Bollengier, T. (1997). Chronological age and Peabody Picture Vocabulary Test performance of persons with mental retardation: New data. *Psychological Reports*, *81*, 1232-1234. doi: 10.2466/pr0.1997.81.3f.1232

Facon, B., & Facon-Bollengier, T. (1999). Chronological age and performance of persons with mental retardation on verbal subtests of the Wechsler Intelligence Scale for Children-Revised, French version. *Psychological Reports*, *85*, 857-862. doi.org/10.2466%2Fpr0.1999.85.3.857

Facon, B., Facon-Bollengier, T., & Grubar, J. C. (2002). Chronological age, receptive vocabulary and syntax comprehension in children and adolescents with mental retardation. *American Journal on Mental Retardation*, *107*, 91-98.

Facon, B., Grubar, J. C., & Gardez, C. (1998). Chronological age and receptive vocabulary of persons with Down syndrome. *Psychological Reports*, *82*, 723-726. doi.org/10.2466%2Fpr0.1998.82.3.723

Facon, B., Magis, D., & Belmont, J. M. (2011). Beyond matching on the mean in developmental disabilities research. *Research in Developmental Disabilities*, *32*, 2134-2147. doi:10.1016/j.ridd.2011.07.029

Facon, B., Magis, D., & Courbois, Y. (2012). On the difficulty of relational concepts among participants with Down syndrome. *Research in Developmental Disabilities*, *33*, 60-68. doi:10.1016/j.ridd.2011.08.014

Facon, B., Magis, D., Nuchadee, M.-L., & De Boeck, P. (2011). Do Raven's Colored Progressive Matrices function in the same way in typical and clinical populations? Insight from the intellectual disability field. *Intelligence*, *39*, 281-291. doi:10.1016/j.intell.2011.04.002

Facon, B., & Nuchadee, M.-L. (2010). An item analysis of Raven's Colored Progressive Matrices among participants with Down syndrome. *Research in Developmental Disabilities*, *31*, 243-249. doi:10.1016/j.ridd.2009.09.011

Facon, B., Nuchadee, M.-L., & Bollengier, T. (2012). A qualitative analysis of general receptive vocabulary of adolescents with Down syndrome. *American Journal on Intellectual and Developmental Disabilities*, *117*, 243-259. doi:10.1352/1944-7558-117.3.243

Farran, E. K., & Jarrold, C. (2003). Visuospatial cognition in Williams syndrome: Reviewing and accounting for the strengths and weaknesses in performance. *Developmental Neuropsychology*, *23*, 173-200. doi:10.1207/S15326942DN231&2_8

Fazio, B. B., Johnston, J. R., & Brandl, L. (1993). Relation between mental age and vocabulary development among children with mild mental retardation. *American Journal on Mental Retardation*, *97*, 541-546.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*, 234–248. doi:10.1111/desc.12019

Fidler, D. J., Daunhauer, L. A., Will, E., Gerlach-McDonald, B., & Schworer, E. (2016). The central role of etiology in science and practice in intellectual disability. *International Review of Research in Developmental Disabilities*, *50*, 1-36. doi:10.1016/bs.irrdd.2016.05.005

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement*, *67*, 565-582. doi:10.1177/0013164406296975

Fisch, G. S., Carpenter, N., Howard-Peebles, P. N., Holden, J. J. A., Tarleton, J., Simensen, R., & Battaglia, A. (2012). Developmental trajectories in syndromes with intellectual disability, with a focus on Wolf-Hirschhorn and its cognitive-behavioral profile. *American Journal on Intellectual and Developmental Disabilities*, *117*, 167-179. doi:10.1352/1944-7558-117.2.167

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, *67*, 373-393. doi:10.1177/0013164406294781

Furlow, C. F., Ross, T. R., & Gagné, P. (2009). The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement*, *33*, 441-464. doi:10.1177/0146621609331959

Gómez-Benito, J., Hidalgo, M. D., & Padilla, J.-L. (2009). Efficacy of effect size measures in logistic regression: An application for detecting DIF. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *5*, 18-25. doi:10.1027/1614-2241.5.1.18

Grela, B. G. (2002). Lexical verb diversity in children with Down syndrome. *Clinical Linguistics & Phonetics*, *16*, 251-263. doi:10.1080/02699200210131987

Griggs, J. L., Sinnayah, P., & Mathai, M. L. (2015). Prader-Willi syndrome: From genetics to behaviour, with special focus on appetite treatments. *Neuroscience and Biobehavioral Reviews*, *59*, 155-172. doi:10.1016/j.neubiorev.2015.10.003

Hansen, A. (2009). Basic conceptual systems (BCSs) – tools for analytic coding, thinking and learning: A concept teaching curriculum in Norway. *Thinking Skills and Creativity*, *4*, 160-169. doi:10.1016/j.tsc.2009.09.001

Hart, B. (1996). The initial growth of expressive vocabulary among children with Down syndrome. *Journal of Early Intervention*, *20*, 211-221. doi.org/10.1177/105381519602000305

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes Publishing Company.

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology*, *49*, 4–14. doi:10.1037/a0027238

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hore, A. P., & Tryon, W. W. (1989). Study of the similar structure hypothesis with mentally retarded adults and nonretarded children of comparable mental age. *American Journal on Mental Retardation*, *94*, 182-188.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*, 329-349. doi.org/10.1207/S15324818AME1404_2

Jonas, R. K., Montojo, C. A., & Bearden, C. E. (2014). The 22q112 deletion syndrome as a window into complex neuropsychiatric disorders over the lifespan. *Biological Psychiatry*, *75*, 351-360. doi:10.1016/j.biopsych.2013.07.019

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, *2*, 389-398. doi:10.1016/S1364-6613(98)01230-3

Karmiloff-Smith, A. (2011). Static snapshots versus dynamic approaches to genes, brain, cognition, and behavior in neurodevelopmental disabilities. *International Review of Research in Developmental Disabilities*, *40*, 1-15. doi:10.1016/B978-0-12-374478-4.00001-0

Karmiloff-Smith, A., Brown, J. H., Grice, S., & Paterson, S. (2003). Dethroning the myth: Cognitive dissociations and innate modularity in Williams syndrome. *Developmental Neuropsychology*, *23*, 229-244. doi:10.1207/S15326942DN231&2_10

Kingdon, D., Cardoso, C., & McGrath, J. J. (2016). Research Review: Executive function deficits in fetal alcohol spectrum disorders and attention-deficit/hyperactivity disorder – a meta-analysis. *Journal of Child Psychology and Psychiatry*, *57*, 116-131. doi:10.1111/jcpp.12451

Knowland, V. C. P., & Thomas, M. S. C. (2011). Developmental trajectories in genetic disorders. *International Review of Research in Developmental Disabilities*, *40*, 43-73. doi:10.1016/B978-0-12-374478-4.00003-4

Krzywinski, M., & Altman, N. (2013). Power and sample size. *Nature Methods*, *10*, 1139-1140. doi:10.1038/nmeth.2738

Lightbody, A., & Reiss, A. (2009). Gene, brain, and behavior relationships in Fragile X syndrome: Evidence from neuroimaging studies. *Developmental Disabilities Research Reviews*, *15*, 343-352. doi:10.1002/ddrr.77

Loveall, S. J., Channell, M. M., Phillips, B. A., Abbeduto, L., & Conners, F. A. (2016). Receptive vocabulary analysis in Down syndrome. *Research in Developmental Disabilities*, *55*, 161-172. doi:10.1016/j.ridd.2016.03.018

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847-862. doi:10.3758/BRM.42.3.847

Martinson, B., & Strauss, A. A. (1941). A method of clinical evaluation of the responses to the Stanford-Binet intelligence test. *American Journal of Mental Deficiency*, *46*, 48-59.

McDonald, J. H. (2014). *Handbook of Biological Statistics* (3rd ed.). Baltimore, MD: Sparky House Publishing.

McKusick-Nathans Institute of Genetic Medicine (2019). OMIM - Online Mendelian Inheritance in Man. Johns Hopkins University, Baltimore, MD. Accessed www.omim.org 28 August 2019.

Mertz, L. G. B., Thaulov, P., Trillingsgaard, A., Christensen, R., Vogel, I., Hertz, J. M., & Østergaard, J. R. (2014). Neurodevelopmental outcome in Angelman syndrome: Genotype-phenotype correlations. *Research in Developmental Disabilities*, *35*, 1742-1747. doi:10.1016/j.ridd.2014.02.018

Mervis, C. B., & John, A. E. (2008). Vocabulary abilities of children with Williams syndrome: Strengths, weaknesses, and relation to visuospatial construction ability. *Journal of Speech, Language, and Hearing Research*, *51*, 967-982. doi:10.1044/1092-4388(2008/071)

Mervis, C. B., Morris, C. A., Bertrand, J., & Robinson, B. F. (1999). Williams syndrome: Findings from an integrated program of research. In H. Tager-Flusberg (Ed.), *Neurodevelopmental disorders* (pp. 65-110). Cambridge, MA: MIT Press.

Meyers, C. E., Dingman, H. F., Attwell, A. A., & Orpet, R. E. (1961). Comparative abilities of normals and retardates of M.A. 6 years on a factor-type test battery. *American Journal of Mental Deficiency*, *66*, 250-258.

Miolo, G., Chapman, R. S., & Sindberg, H. (2005). Sentence comprehension in adolescents with Down syndrome and typically developing children: Role of sentence voice, visual context, and auditory-verbal short-term memory. *Journal of Speech, Language, and Hearing Research*, *48*, 172-188. doi:10.1044/1092-4388(2005/013)

Miolo, G., Chapman, R. S., & Sindberg, H. (2005). Sentence comprehension in adolescents with Down syndrome and typically developing children: Role of sentence voice, visual context, and auditory-verbal short-term memory. *Journal of Speech, Language, and Hearing Research*, *48*, 172-188. doi: 10.1044/1092-4388(2005/013)

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … Chartier, C. R. (2018). The Psychological Science Accelerator□: Advancing Psychology Through a Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science*, *1*, 501-515. doi.org/10.1177/2515245918797607

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*, 257-274. doi:10.1177/014662169602000306

O'Hearn, K., & Landau, B. (2007). Mathematical skill in individuals with Williams syndrome: Evidence from a standardized mathematics battery. *Brain and Cognition*, *64*, 238-246. doi:10.1016/j.bandc.2007.03.005

Osterlind, S. J. (1983). *Test item bias*. Newbury Park, CA: Sage Publications.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning (2nd ed.)*. Thousand Oaks, CA: Sage Publications.

Pei, L. K., & Li, J. (2010). Effects of unequal ability variances on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT likelihood ratio for DIF detection. *Applied Psychological Measurement*, *34*, 453-456. doi:10.1177/0146621610367789

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharray (Eds.), *Handbook of Statistics: Psychometrics* (Vol. 26, pp. 125-167). Amsterdam, NL: Elsevier.

Phillips, C. E., Jarrold, C., Baddeley, A., Grant, J., & Karmiloff-Smith, A. (2004). Comprehension of spatial language terms in Williams syndrome: Evidence for an interaction between domains of strength and weakness. *Cortex*, *40*, 85□101. doi:10.1016/S0010-9452(08)70922-5

Phillips, B. A., Loveall, S. J., Channell, M. M., & Conners, F. A. (2014). Matching variables for research involving youth with Down syndrome: Leiter-R versus PPVT-4. *Research in Developmental Disabilities*, *35*, 429-438. doi:10.1016/j.ridd.2013.11.016

Polišenská, K., & Kapalková, S. (2014). Language profiles in children with Down syndrome and children with language impairment: Implications for early intervention. *Research in Developmental Disabilities*, *35*, 373-382. doi:10.1016/j.ridd.2013.11.022

Polišenská, K., Kapalková, S., & Novotková, M. (2018). Receptive language skills in Slovak-speaking children with Intellectual Disability: Understanding words, sentences and stories. *Journal of Speech, Language, and Hearing Research*, *61*, 1731-1742. doi: 10.1044/2018_JSLHR-L-17-0029

Price, J., Roberts, J., Vandergrift, N., & Martin, G. (2007). Language comprehension in boys with fragile X syndrome and boys with Down syndrome. *Journal of Intellectual Disability Research*, *51*, 318-326. doi:10.1111/j.1365-2788.2006.00881.x

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raven, J. C., Court, J. H., & Raven, J. (1998). *Progressive Matrices Couleur [Colored Progressive Matrices]*. Paris: Les Editions du Centre de Psychologie Appliquée.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*, 193-203. doi:10.1111/j.1745-3984.1988.tb00302.x

Richardson, F. M., & Thomas, M. S. C. (2009). Language development in genetic disorders. In E. Bavin (Ed.), *The Cambridge handbook of child language* (pp. 459-471). Cambridge, UK: Cambridge University Press.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*, 105-116. doi:10.1177/014662169301700201

Santucci, H., & Hélal, A. (1969). Les caractères spécifiques du préadolescent débile à l'épreuve du Binet-Simon. In R. Zazzo (Ed.), *Les débilités mentales* (pp. 288-316). Paris: Armand Colin.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flawed items in the test adaptation process. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*, 170-187. doi:10.1080/13803611.2013.767621

Somerville, M. J., Mervis, C. B., Young, E. J., Seo, E.-J., del Campo, M., Bamforth, S., Peregrine, E., Loo, W., Lilley, M., Pérez-Jurado, L. A., Morris, C. A., Scherer, S. W., & Osborne, L. R. (2005). Severe Expressive-Language Delay Related to Duplication of the Williams–Beuren Locus. *New England Journal of Medicine*, *353*(16), 1694-1701. doi.org/10.1056/NEJMoa051962

Spitz, H. H. (1982). Intellectual extremes, mental age, and the nature of human intelligence. *Merrill-Palmer Quarterly of Behavior and Development*, *28*, 167-192.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370. doi:10.1111/j.1745-3984.1990.tb00754.x

Thomas, M. S. C. (2016). Understanding delay in developmental disorders. *Child Development Perspectives*, *10*, 73-80. doi:10.1111/cdep.12169

Thomas, M. S. C., Annaz, D., Ansari, D., Serif, G., Jarrold, C., & Karmiloff-Smith, A. (2009). Using developmental trajectories to understand developmental disorders. *Journal of Speech, Language, and Hearing Research*, *52*, 336-358. doi:10.1044/1092-4388(2009/07-0144)

Thomas, M. S. C., Purser, H. R., & van Herwegen, J. (2011). Cognition: The developmental trajectories approach. In E. K. Farran & A. Karmiloff-Smith (Eds.), N*eurodevelopmental disorders across the lifespan: A neuroconstructivist approach* (pp. 13-35). Oxford, UK: Oxford University Press.

Warren, S., Brady, N., Sterling, A., Fleming, K., & Marquis, J. (2010). Maternal responsivity predicts language development in young children with fragile X syndrome. *American Journal on Intellectual and Developmental Disabilities*, *115*, 54-75. doi:10.1352/1944-7558-115.1.54

Whitmore, M. L., & Schumacker, R. E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement*, *59*, 910-927. doi:10.1177/00131649921970251

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, CAN: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Prince George, CAN: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.

Table 1: Descriptive statistics for RAVEN, BOEHM, chronological age and gender of participants with or without ID. Cronbach's alpha coefficients for the RAVEN and the BOEHM are also given.

	RAVEN*				BOEHM				Chronological age			Gender	
	Mean	S.D.	Min - max	Cronbach's α	Mean	S.D.	Min - max	Cronbach's α	Mean	S.D.	Min - max	M	F
Participants without ID	15.61	4.51	4 - 30	.721	50.24	12.59	17 - 72	.938	4.57	0.80	2.55 - 6.44	280	277
Participants with ID	15.61	4.51	4 - 30	.719	50.67	13.19	14 - 72	.941	12.20	3.18	4.69 – 21.85	320	237

Note. *Matching variable. N=557 for each group. RAVEN = Raven Colored Progressive Matrices; BOEHM = Test des Concepts de Base, F = females; M = males.

Table 2: Correlation coefficients between chronological age, test scores and gender of participants with and without intellectual disability.

	Participants without ID				Participants with ID			
	CA	RAVEN	BOEHM	Gender	CA	RAVEN	BOEHM	Gender
CA[a]	-	.562**	.724**	-.019	-	.097*	.104*	.015
RAVEN[a]		-	.655**	.010		-	.631**	-.078
BOEHM[a]			-	.009			-	-.049
Gender[b]				-				-

[a]Pearson's product moment correlation; [b]Point-biserial correlation; **$p < .000001$; *$p < .05$; CA = chronological age.

A: no DIF **B: uniform DIF** **C: nonuniform DIF**

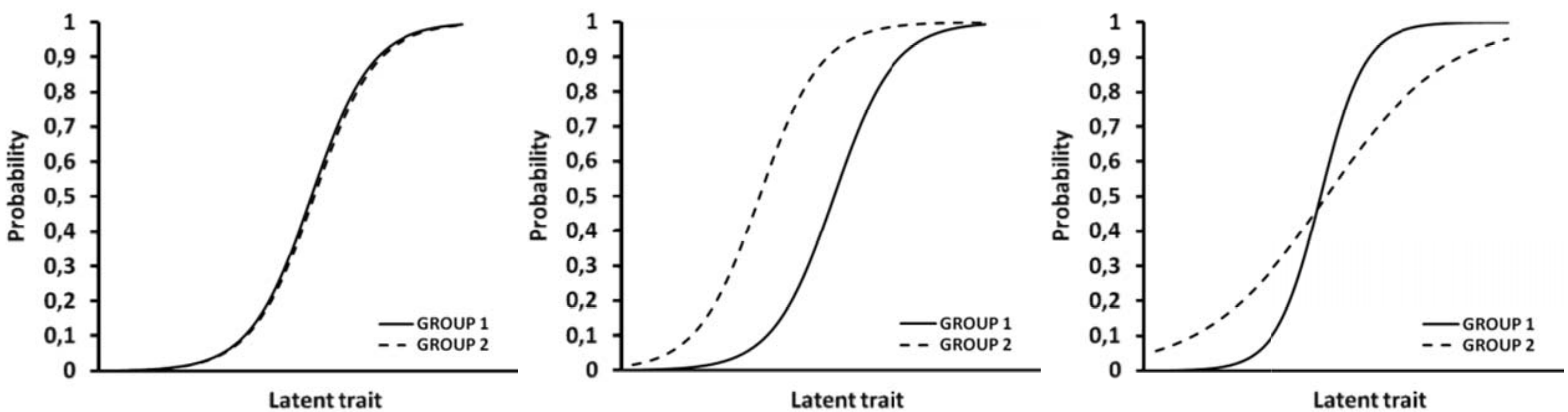

Figure 1: Three hypothetical examples of item characteristic curves for two groups of examinees.

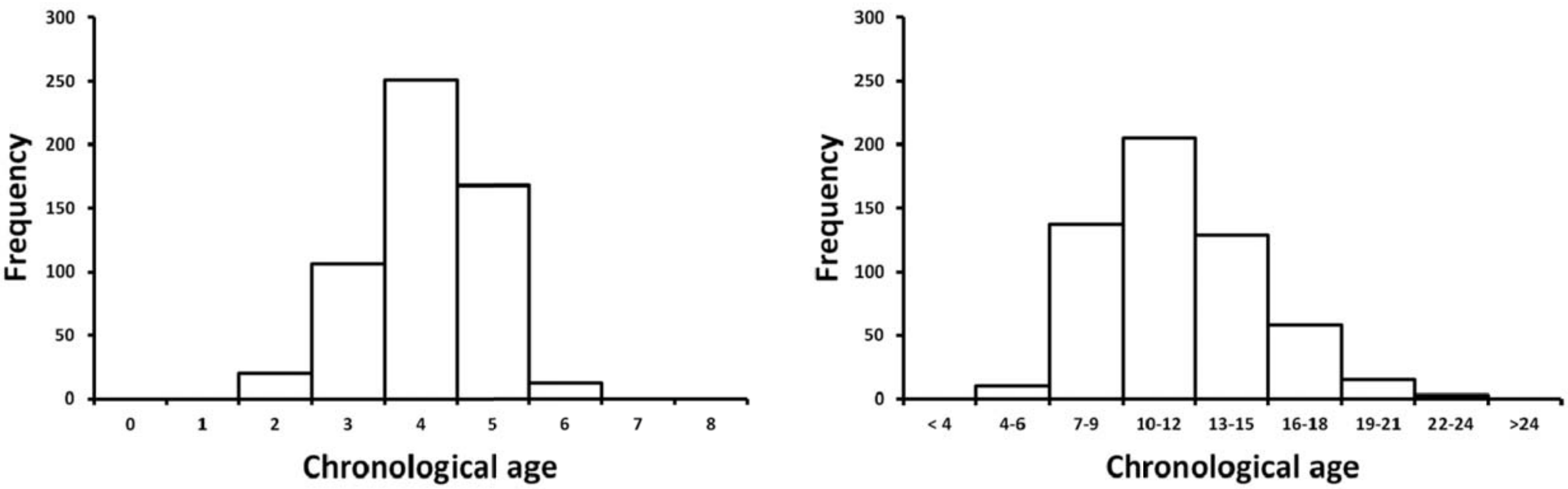

Figure 2: Distribution of chronological ages (in years) of TD participants (left panel) and of participants with ID (right panel).

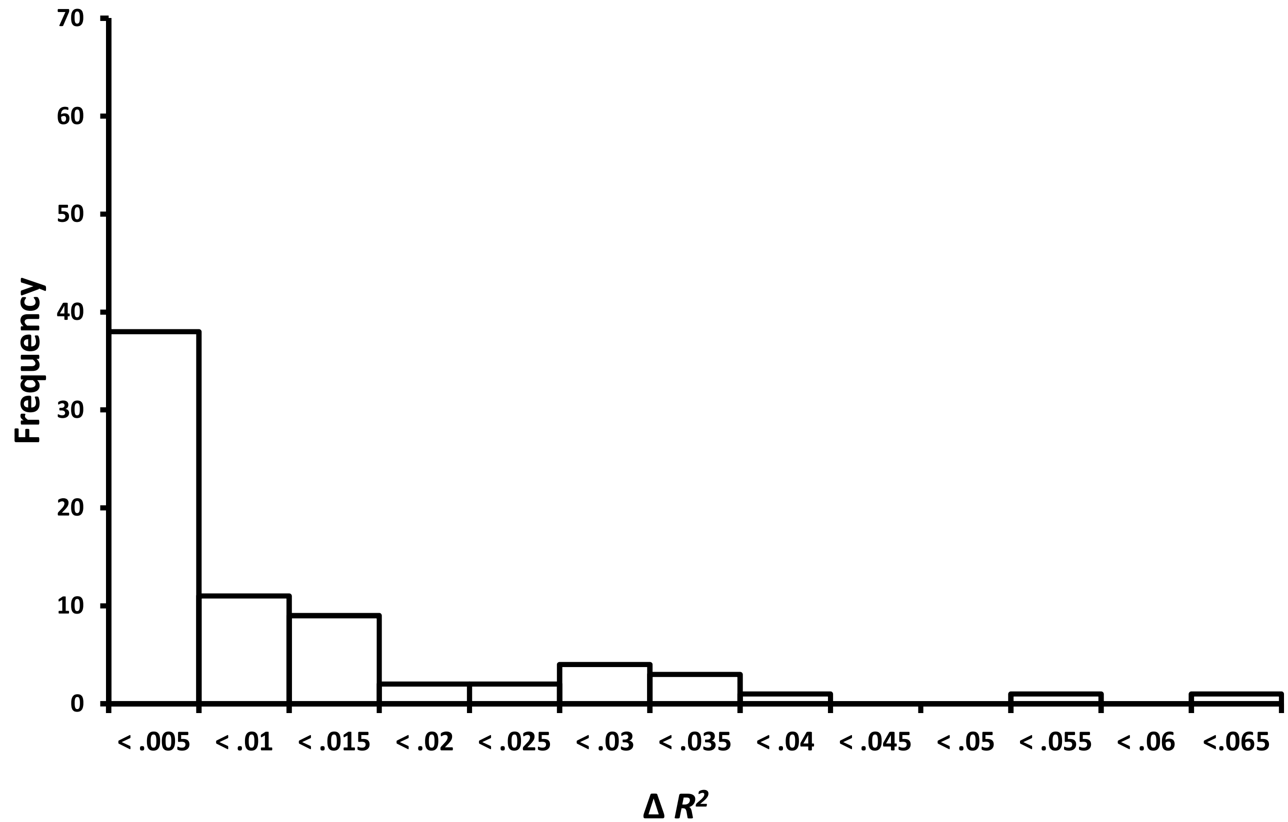

Figure 3: Distribution of ΔR^2 for the status variable. According to Jodoin and Gierl (2001), the effect size of a variable can be considered *negligible* if $\Delta R^2 < .035$, *moderate* if $.035 \leq \Delta R^2 < .07$, and *large* if $\Delta R^2 \geq .07$.

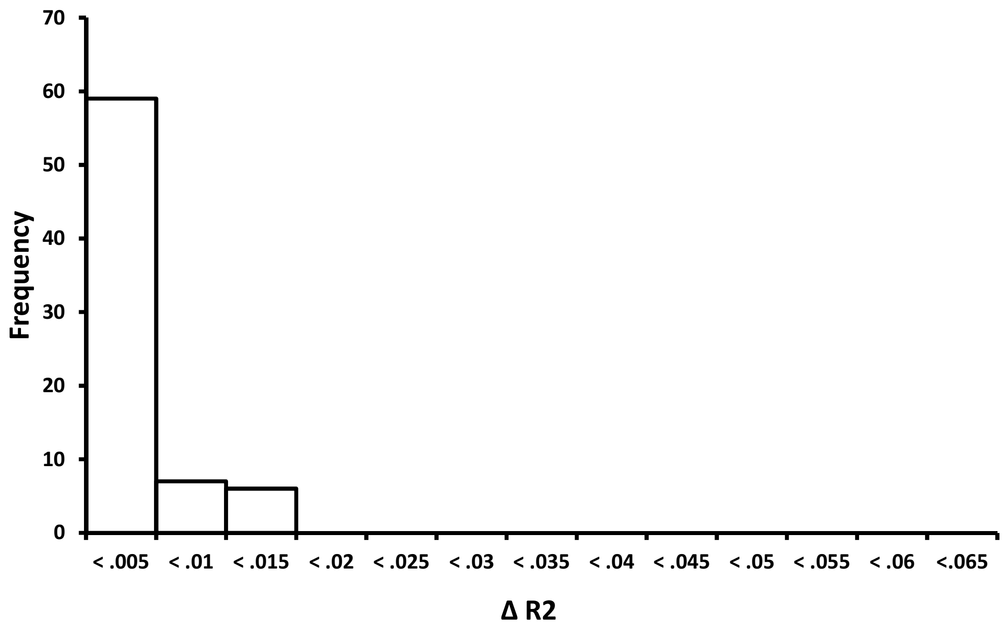

Figure 4: Distribution of ΔR^2 for the RAVEN × status interaction. According to Jodoin and Gierl (2001), the effect size of a variable can be considered as *negligible* if $\Delta R^2 < .035$, *moderate* if $.035 \leq \Delta R^2 < .07$, and *large* if $\Delta R^2 \geq .07$.

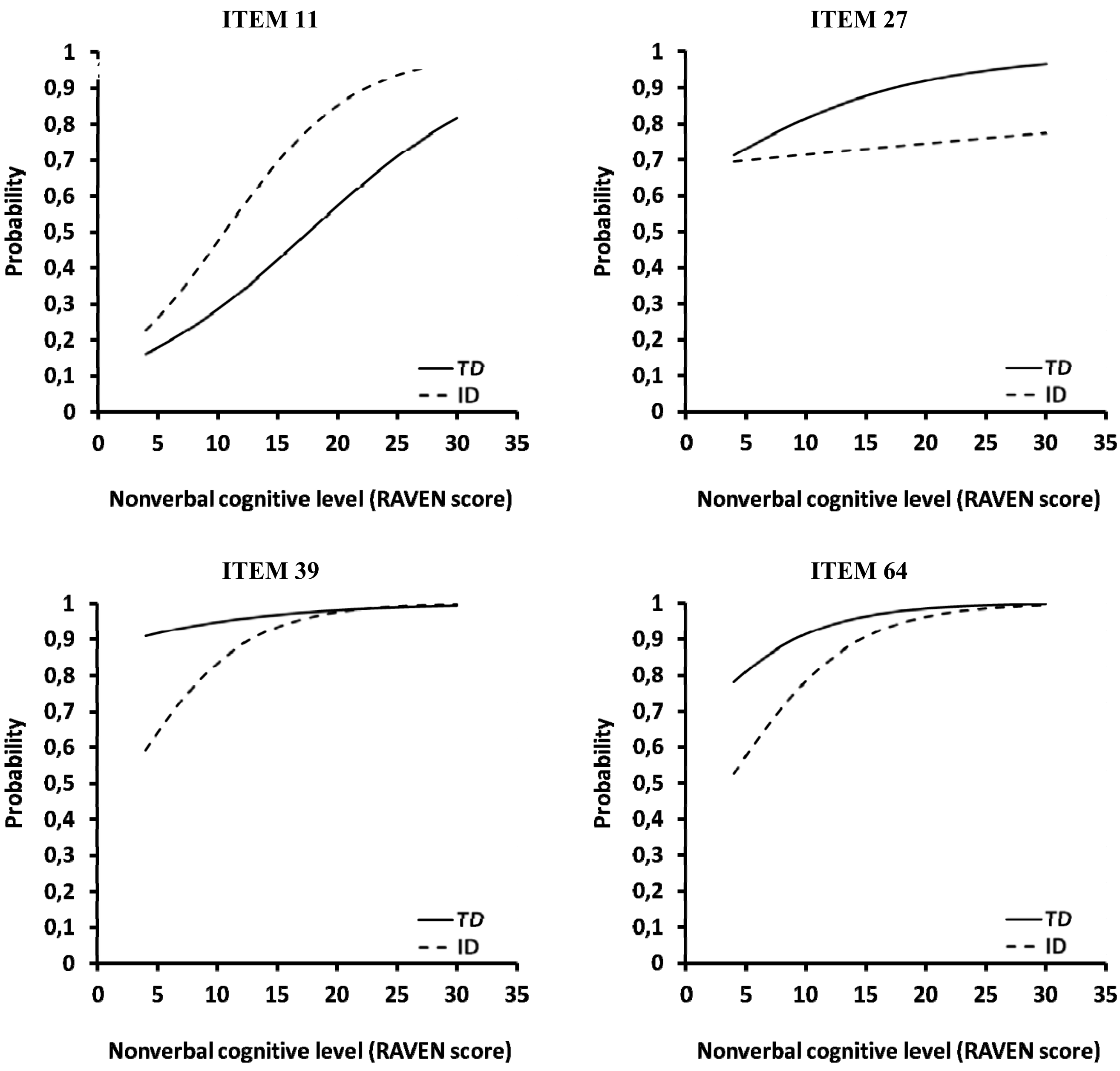

Figure 5: Logistic curves of the four items flagged as showing DIF. Solid lines represent the reference group (TD participants), dashed lines the focal group (participants with ID).

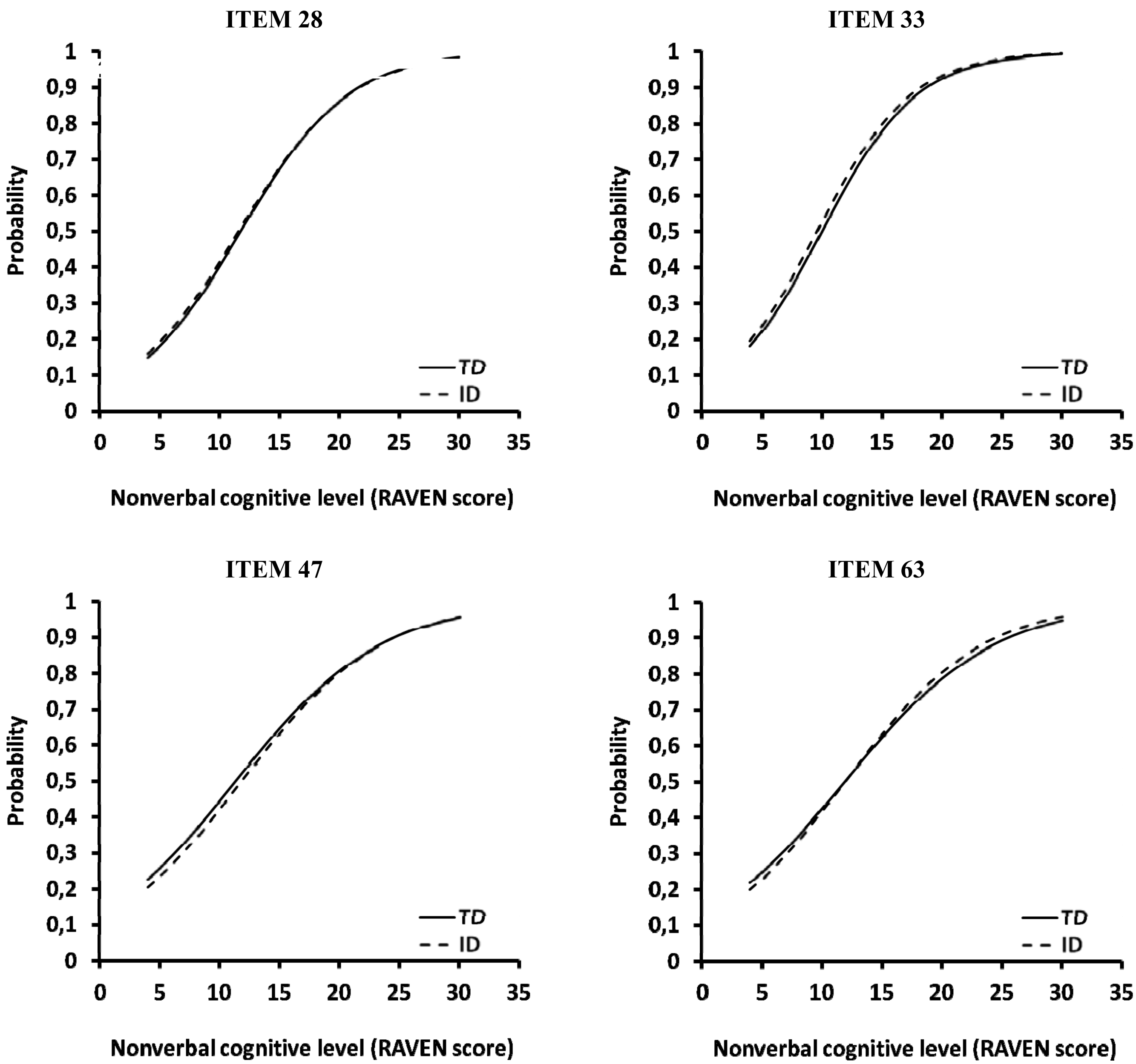

Figure 6: Four examples of logistic curves almost indistinguishable. Solid lines represent the reference group (TD participants), dashed lines the focal group (participants with ID).

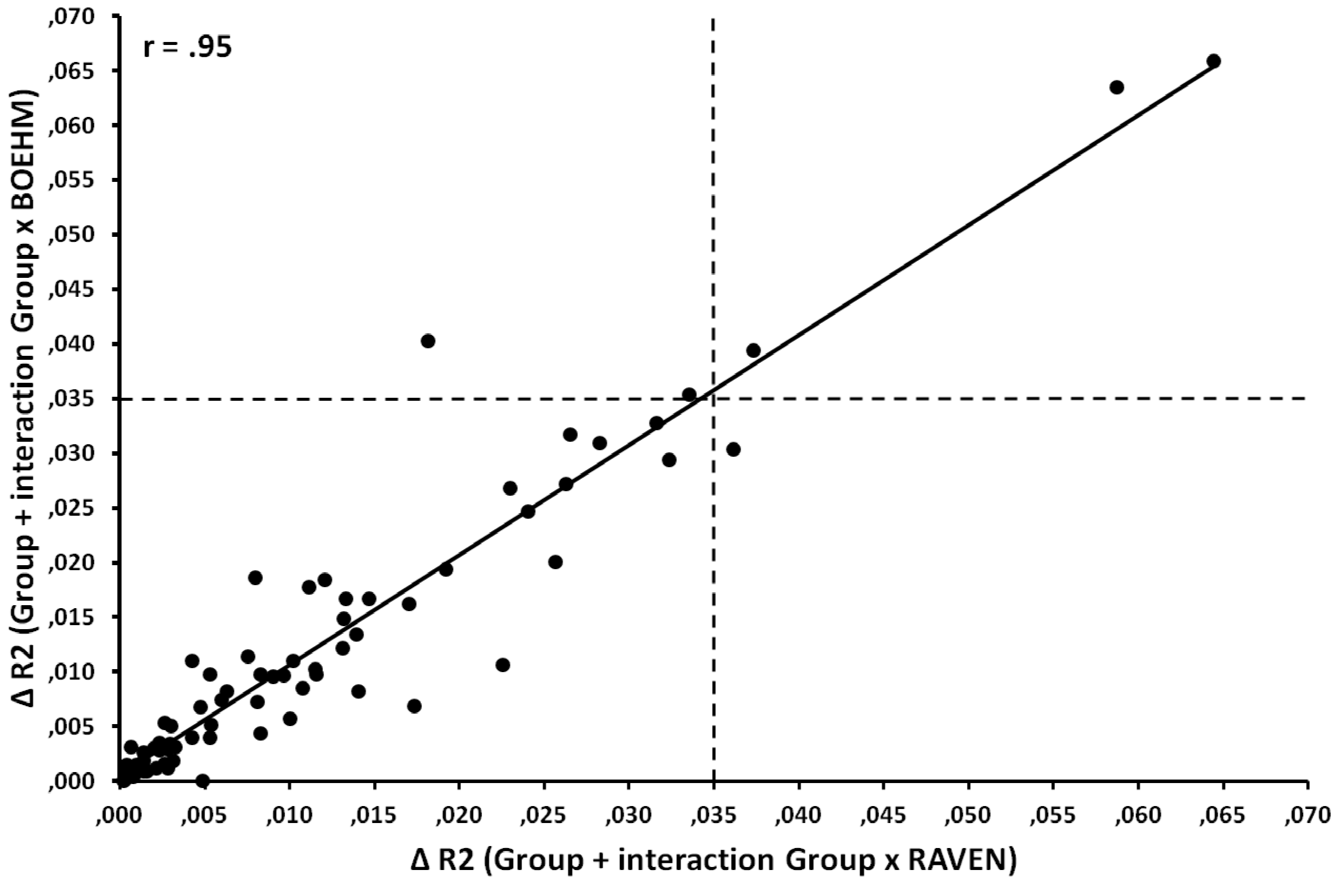

Figure 7: Bivariate distribution of ΔR^2 from the two successive analyses. Solid line represents the regression line, dashed lines the Jodoin and Gierl (2001) threshold of moderate effect size.

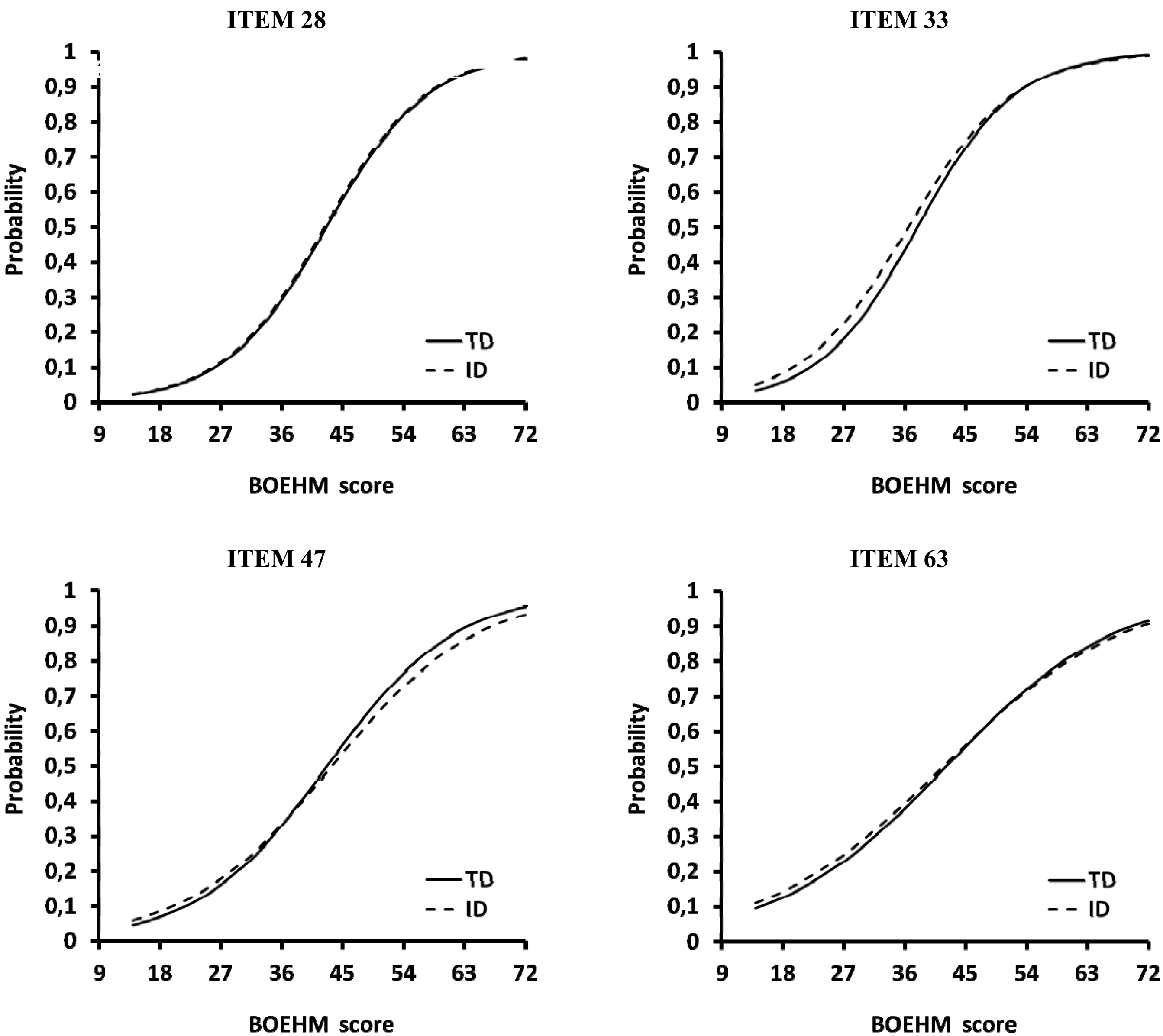

Figure 8: Logistic curve of items 28, 33, 47 and 63. Note that the x-axis is now the BOEHM score and that these are the same items as in Figure 5. Solid lines represent the reference group (TD participants), dashed lines the focal group (participants with ID).

Appendix

Percentages of correct responses on each item of the BOEHM test for participants with and without ID matched on RAVEN score. Items' rank orders of difficulty are also shown.

Items	% of correct responses		Items' rank order	
	TD	**ID**	**TD**	**ID**
1 Show me the cat that's on ***top*** of the car.	98	97	1	1
2 Show me the rabbit whose ears point ***downwards***.	85	80	27	27
3 Look at the food bowls. Show me the one that's ***empty***.	94	95	10	4
4 Show me the ball that's ***under*** the table.	73	76	34	30
5 Show me the balloon that's ***highest***.	96	92	6	11
6 Show me the clown who ***lacks*** a hat.	96	93	9	8
7 Show me the animal that's ***next*** to the rabbit.	88	80	24	26
8 Look at this apple (the examiner indicates the apple up above, then the row of fruits below). Show me the ***other*** apple.	96	95	7	6
9 Show me the dog that is going ***up***.	88	80	23	28
10 Show me the jar that is ***full*** of candies.	93	90	16	13
11 Show me the toys that are ***outside*** the box.	44	69	64	39
12 Show me ***all*** the shoes.	94	89	13	17
13 Here is a tree (the examiner indicates the tree). Show me the dog that is ***nearest*** to the tree.	93	88	17	18
14 These girls are eating pasta. Show me the girl who has ***finished***.	98	94	2	7
15 Show me the ***smallest*** fish.	93	87	14	19
16 Show me the child who is ***crossing*** the bridge.	73	73	33	36
17 Show me the shoe that is ***different***.	81	81	30	25
18 Show me the ***longest*** train.	87	85	26	22
19 Show me the cat that's ***in front of*** the chair.	72	73	35	34
20 Show me the ***pair*** of horses.	29	42	71	69
21 Show me the box with the crayons ***around*** it.	75	76	32	31
22 Show me the ***biggest*** tree.	97	96	3	2
23 Look at these children. (the examiner indicates the children in the pictures). Show me the picture with ***many*** children.	94	93	11	9
24 Look at this mouse (the examiner indicates the mouse above, then those below and says:) Show me the mouse that's the ***same***.	90	90	20	15
25 Show me the box that has ***the most*** crayons.	92	91	18	12
26 Show me the ***biggest*** ice cream.	97	96	4	3
27 Show me the girl who is ***in front of*** the boy in the line.	88	73	25	33
28 Look at this box (the examiner indicates the box). Show me the ball that is ***farthest*** from the box.	67	68	38	42
29 Show me the child who is ***the last*** in line.	70	73	36	32
30 Look at the dogs' tails. Show me the dog with the ***shortest*** tail.	40	45	66	65
31 Show me the child who is ***the last*** in line.	58	66	48	45
32 Look at these blocks. Show me the block that is ***lowest***.	93	90	15	14
33 Show me the ducks that are swimming ***together***.	76	77	31	29
34 Show me the box with the ***fewest*** fries.	34	43	69	68
35 Show me the mouse that's in ***the middle***.	66	68	39	41
36 Show me ***the first*** car in the line.	84	81	28	23

37 Show me the picture where the bear is ***between*** the blocks.	64	66	42	44
38 Show me the jar with ***the least*** jam.	53	54	52	55
39 Look at the dogs who are playing. Show me the dog that is going ***through*** the hoop.	97	92	5	10
40 Look at the cakes. Show me the cake that is ***whole***.	59	64	46	46
41 Look at the t-shirts. Show me the t-shirt that is ***middle-sized***.	69	59	37	52
42 Look at the box, the feather, the wheel, and the ball. Show me what is ***always*** found on a bike.	94	95	12	5
43 Look at the frogs on the log. Show me the frog that is ***on the right***.	57	51	50	60
44 Look at the eggs that have fallen out of the basket. Show me the egg that is ***in the corner*** of the table.	55	58	51	53
45 Look at the pictures of the boy. Show me the picture of the boy ***before*** he hurt his finger.	48	52	59	59
46 Look at the train with the animals. Show me the animal that is at ***the end*** of the train.	64	69	43	40
47 Look at the bees and the flower. Show me the bee that is ***furthest*** from the flower.	65	64	40	49
48 Look at the pictures of a plane. Show me the picture that shows only a ***part of*** the plane.	52	60	53	51
49 Look at the marbles. Show me the marble that is in the ***center*** of the square.	39	43	67	67
50 Look at the plants. Show me the plant that has ***a few*** flowers but ***not a lot***.	50	47	56	64
51 Look at the house with the windows. Show me the window that is ***over*** the door.	89	81	21	24
52 Look at the pictures of blocks. Show me the picture where the blocks are ***separated***.	51	62	54	50
53 Look at the pictures of dogs. Show me the picture where all the dogs are ***in a row***.	59	68	47	43
54 Look at the rabbit, the cat, the pig and the dog. Show me the animal that is ***near*** the rabbit.	84	72	29	38
55 Look at the pictures of dogs. Show me the picture where there are ***a few*** dogs.	45	50	62	61
56 Look at the pictures of flowers and bees. Show me the picture where there is one bee on ***each*** flower.	50	44	55	66
57 Look at the apple, the grape, the pear, and the chair. Show me the one that a child ***never*** eats.	38	48	68	62
58 Look at the children. Show me the child who is ***beginning*** to climb.	41	37	65	70
59 Look at the house and the cars. Show me the ***second*** car from the house.	46	56	61	54
60 Look at the windows. Show me the ***widest*** window.	57	64	49	47
61 Look at the blocks and the child. Show me the block that is ***farthest*** from the child.	89	85	22	21
62 Look at the pies. Show me the picture with ***half*** a pie.	63	72	45	37
63 Look at the pictures of t-shirts and shorts. Show me the picture where the t-shirt and the shorts go ***together***.	63	64	44	48
64 Look at the children and the rope. Show me the child who is jumping ***over*** the rope.	96	89	8	16
65 Look at the candies in the drawer. Show me the picture that has as ***many*** candies as are in the drawer.	64	73	41	35
66 Look at the line of animals. Show me the animal that is ***third*** in line.	46	53	60	57
67 Look at the chair and the toys. Show me the toy that is ***behind*** the chair.	92	86	19	20
68 Look at the dinosaurs. Show me the dinosaur that is ***on the left***.	50	48	57	63
69 Look at the rocket in the circle (examiner points to it). ***Skip*** one rocket and show me the ***next*** one.	14	14	72	72
70 Look at the pictures of the girl. Show me the girl who is leaning ***forward***.	50	54	58	56
71 Look at the rabbit in the square. Show me the square where it would be if it hopped ***backwards***.	45	53	63	58
72 Look at these hands that are holding candies. Show me the picture where the hands have an ***equal*** number of candies.	33	34	70	71