Psicologia: Teoria e Pesquisa
Mai-Ago 2004, Vol. 20 n. 2, pp. 103-111

Cultural Bias in the SON-R Test: Comparative Study of Brazilian and Dutch Children[1]

Peter J. Tellegen[2]
University of Groningen
Jacob A. Laros
Universidade de Brasília

ABSTRACT – The present study, including 83 Brazilian and 51 Dutch children, evaluated the presence of cultural bias in items of the SON-R 5½-17 that make use of concrete objects and situations. Two procedures were followed to detect item bias. The first consisted of asking the children, immediately after an incorrect answer, whether they recognized the pictures. The second procedure compared item difficulties of the Brazilian children with those of the Dutch children belonging to the standardization sample of the SON-R 5½-17. Fourteen items were detected with bias: ten of these favored the Dutch group and four the Brazilian group. The cultural disadvantage for Brazilian children is rather small, taking the large amount of investigated items into account. This study indicated which items of the SON-R 5½-17 should be improved, not only for reasons of cultural bias, but also because children, irrespective of their cultural background, encountered problems with the recognition of several pictures.

Key words: cultural bias; item bias; nonverbal intelligence test.

Viés Cultural no Teste SON-R: Estudo Comparativo entre Crianças Brasileiras e Holandesas

RESUMO – No presente estudo, incluindo 83 crianças brasileiras e 51 holandesas, verificou-se a presença de viés cultural nos itens do SON-R 5½-17 que usam objetos e situações concretas. Dois procedimentos foram seguidos para detectar viés do item. No primeiro, perguntou-se às crianças, imediatamente depois de uma resposta errada, se elas reconheceram os desenhos utilizados nos itens. No segundo procedimento, comparou-se a dificuldade dos itens para as crianças brasileiras com a dificuldade dos itens para as crianças holandesas da amostra de normatização do SON-R 5½-17. Identificaram-se quatorze itens com viés, dos quais dez favorecem as crianças holandesas e quatro as crianças brasileiras. A desvantagem cultural para as crianças brasileiras é bastante pequena, levando em consideração o grande número de itens investigado. Este estudo indicou quais itens do SON-R 5½-17 precisam ser melhorados, não só por razões de viés cultural, mas também porque crianças, independentemente do *background* cultural, encontraram problemas com o reconhecimento de vários desenhos.

Palavras-chave: viés cultural; viés do item; teste não-verbal de inteligência.

In Brazil and other South American countries there is a great need for standardized and validated psychological tests, especially for tests in relation to intelligence for children and youth (Hu & Oakland, 1991; Oakland, Wechsler, Bensuan & Stafford, 1994; Muñiz, Prieto, Almeida & Bartram, 1999). A nonverbal intelligence test that might fill this need is the SON-R 5½-17, the Snijders-Oomen nonverbal intelligence test for children and adolescents in the ages of 5½ to 17 years (Snijders, Tellegen & Laros, 1989).

The SON-R 5½-17 is an individual intelligence test for general application, which can be administered without the use of spoken or written language. The first SON-test was published in the Netherlands 1943 by Mrs. Nan Snijders-Oomen as a result of her work at an institute for deaf children. The test was originally developed in order to be able to assess the learning ability of children who were severely handicapped in their language development. At that time, existing nonverbal intelligence tests were not suited for the examination of a broad spectrum of learning abilities because they consisted mainly of performance tests related to spatial abilities (like form boards, mazes, and mosaics). The first SON-test of 1943 consisted of nonverbal subtests related to abstract and concrete reasoning, and contained norms for deaf children from 4 to 14 years of age. At present, the fourth generation of SON nonverbal intelligence tests exist of two versions: one for younger children, the SON-R 2½-7, and one for older children, the SON-R 5½-17. The present article discribes a study which was realized with the SON-test for older children.

The SON-R 5½-17 consists of the following subtests: Categories, Analogies, Situations, Stories, Patterns, and Hidden Pictures. The first three are multiple choice tests, the remaining four are action tests. In action tests the solution has to be sought in an active manner, which makes observation of behavior possible. The SON-R 5½-17 can be divided into four

1 Estudos realizados com apoio da CAPES, CNPq e do Fundo SON da Universidade de Groningen, Holanda. Agradecimentos aos estudantes da Universidade de Brasília e da Universidade de Groningen.

2 Endereço: Institute of Psychology Heymans, Grote Kruisstraat 2/1, Groningen, The Netherlands, 9712 TS. E-mail: P.J.Tellegen@ppsw.rug.nl

types of tests according to their contents: abstract reasoning tests (Categories and Analogies), concrete reasoning tests (Situations and Stories), spatial tests (Mosaics and Patterns), and perceptual tests (Hidden Pictures). The subtest items are presented using an adaptive test procedure. Adaptive procedures are aimed at limiting the number of items to be administered with relatively little loss of reliability (Weiss, 1982). The adaptive procedure of the SON-R 5½-17 is based on the division of its subtests in two or three parallel series of about 10 items. The first series of a subtest serves to estimate the subject's general level of performance. Of the following series only those items that can improve and refine this first estimation are administered. The adaptive test procedure reduces the amount of items to be presented by about 50%. The application of the test takes about 90 minutes; the shortened version, consisting of four subtests, takes approximately 45 minutes.

The standardization was performed using a representative sample of children of the Netherlands, consisting of 1,350 subjects from 6 to 14 years of age. Each age group was represented by a sample of 150 subjects which was stratified according to sex, educational type, and demographic variables. The expansion of the norms to the ages of 5½ to 17 was achieved through extrapolation. Norm tables for 38 age groups make it possible to draw comparisons at the subtest level. The total test result is represented as an IQ score (with probability interval), as a percentile score, and as a reference age. In addition to the 1,350 hearing subjects, also 768 deaf children were examined with the test. A computer program for the calculation of standardized scores is supplied with the test. After the birth date, the test date, and the raw subtest scores have been entered, the computer program calculates the standardized scores automatically based on the subject's exact age. The SON-R 5½-17 has been reviewed by the COTAN, the test commission of the Netherlands Institute of Psychologists, and received the highest possible ratings on all seven categories of evaluation. The categories of reviewing were as follows: (1) basics of the construction of the test; (2) quality of the test materials; (3) quality of the manual; (4) norms; (5) reliability; (6) construct validity, and (7) criterion validity. The SON-R 5½-17 is being used in various countries; the manual is available in English, German, and Dutch.

Advocates of culture fair intelligence tests have criticized traditional tests for general intelligence, like the Wechsler intelligence tests and the Standford-Binet tests, because they often make an appeal to specific language skills, and in so doing place members of cultural minority groups at a disadvantage. Advocates of learning potential tests have critized traditional tests for general intelligence because these tests would measure the end result of prior learning, rather than learning potential (Tellegen & Laros, 1993). By focussing on the end result of prior learning, these tests would underestimate the learning ability of persons from lower socio-economic background, and of members of ethnic minorities. One could state that these tests focus more on "crystallized intelligence" rather than on "fluid intelligence"(Cattell, 1971).

The SON-R 5½-17 differs in three essential aspects from traditional intelligence tests: in the first place, it does not require specific language abilities; in the second place, it uses an adaptive test procedure; and in the third place, it offers feedback after each item which informs the subject whether the answer is correct. A major advantage of giving feedback is that the subject is giving the opportunity to learn during the test administration. With these essential aspects, the SON-R 5½-17 shows more resemblances to culture faire intelligence tests and tests for learning potential than to traditional tests for general intelligence.

In order to contribute to the demand in Brazil for intelligence tests of good quality, the SON-R 5½-17 has to be standardized and validated for that country. Prior to standardization, however, it is necessary to verify whether the materials used in the test are familiar to Brazilian children and adolescents. To obtain such evidence the present study was undertaken. Thus, the goal of the present study is to discern if, and to what extent, adaptations of the test materials of the SON-R 5½-17 are required in order to assess the construct of (nonverbal) intelligence in Brazil with this test in a fair way. This goal is in accordance with the guidelines on test use of the International Test Commission (Van de Vijver & Hambleton, 1996). One of the guidelines states that test developers/publishers should provide evidence that item content and stimulus materials are familiar to all intended populations.

The fact that the items and the examples of the SON-R do not need to be translated makes the test potentially suitable for international and cross-cultural research. The adaptation process of nonverbal tests for multiple cultures does not include the difficult and often extremely problematic test translation phase and is therefore much less complicated than for (partly) verbal tests.

The research finding that immigrant children in the Netherlands (mainly children from Morocco, Turkey, Suriname and the Dutch Antilles) perform better on the SON-tests than on traditional intelligence tests like the WISC-R (Laros & Tellegen, 1991; Tellegen, Winkel, Wijnberg-Williams & Laros, 1998) is an indication of the culture-fairness of the SON-R for immigrant groups in the Dutch society.

One of the reasons why immigrant children attain relatively lower mean scores on traditional intelligence tests is the strong emphasis of these tests on verbal abilities and specific knowledge learned in school. This is especially the case with the so-called omnibus intelligence tests like the Wechsler scales that contain subtests like Information and Vocabulary (Helms-Lorenz & Van de Vijver, 1995). The fact that minority groups show lower mean scores on a test, however, does not necessarily mean that the test is culturally biased. Van de Vijver and Poortinga (1992) argue that the desirability of cultural loadings in measurement procedures is determined by the intention of the test in question. If a particular test is intended to measure knowledge gained during a course at school it is to be expected that culture-specific knowledge will be assessed. In that case, cultural loadings are unavoidable and even desirable. In general, a distinction can be made between generalizations about achievements and about aptitudes. In the latter case, cultural loadings are undesirable (Helms-Lorenz & Van de Vijver, 1995).

A second research result with positive implications for the culture-fairness of the SON-tests for immigrants is the finding that there is no relation between length of stay in the Netherlands and their IQ-scores, suggesting that in the SON test, intelligence is not dependent on knowledge of the Dutch language (Snijders et al., 1989). As a third research result we should mention that the performance of immigrant children is similar on the SON-R subtests with meaningful pictures compared to the subtests that use materials of an abstract nature.

The aforementioned positive indications of the culture-fairness of the SON-tests are based on research results with immigrant groups in the Netherlands and offer no guarantee for the culture-fairness of the test in a South-American country. The present study was undertaken to obtain indications of the degree of culture-fairness of the SON-R in Brazil.

Method

Participants

The Brazilian sample included 83 children (41 male, 42 female) ranging in age from 7 to 14 years (M = 10.5, *SD* = 2.1). The children were recruited from two state schools in Brasilia. Within the schools children were selected on basis of their age; the children who were selected had their birthday as close as possible half a year from the test date. The Dutch sample consisted of 51 children (24 male, 27 female) ranging in age from 7 to 12 years ($M = 9.9$, $SD = 1.3$). The participants were recruited from three schools in the northern part of the Netherlands. The same selection criteria were used as for the Brazilian sample.

Instruments

The SON-R 5½-17 is the revised version of the Snijders-Oomen Nonverbal intelligence test for children and adolescents of 5½ to 17 years (Snijders et al., 1989; Tellegen & Laros, 1993). The test consists of seven subtests, which are, in order of administration: Categories, Mosaics, Hidden Pictures, Patterns, Situations, Analogies, and Stories. The standardization of the SON-R 5½-17 in the Netherlands is based on a nationwide sample of 1,350 children and adolescents varying in age from 6 to 14 years. The reliability coefficient (alpha stratified) of the IQ increases from .90 at six years to .94 at fourteen years with a mean value of .93. The average reliability of the subtests is .76. The validity of the SON-R 5½-17 is evident from the clear relationship with different indicators of school career such as school type, class repetition and school report marks. The multiple correlation of the SON-R IQ with these indicators of school career is .59.

The subtests of the SON-R 5½-17 can be divided in two types of tests according to the material that is being used: tests that use meaningful picture material (Categories, Situations, and Stories) and tests that use non-meaningful materials such as geometrical forms (Mosaics, Patterns, and Analogies). Hidden Pictures is a case on its own because the task in this subtest, recognition, is independent of the type of material used. In the present study only subtests that use meaningful picture material were included, because cultural bias is more likely to occur with this kind of subtests than with those that use non-meaningful materials such as geometrical forms (Jensen, 1980). In the subtest Categories, a child has to choose two pictures that are missing in a certain category out of five possible pictures. The task in Situations is to indicate the missing parts of drawings of concrete situations. In Stories the child has to order a number of cards in such a way that they form a logical story. Categories and Situations are multiple choice tests, while Stories is a so-called "action" test, where the child has to construct the solution rather than to choose the right alternative. The subtest Categories consists of 27 items, Situations of 33 items, and Stories of 20 items.

Procedure

The first step in this study was the translation of the instructions into the Portuguese language. After obtaining parental permission, the SON-R subtests Categories, Situations and Stories were administered to the Brazilian children. Six graduate psychology students administered the subtests after being trained in the administration of these subtests. Supervision was provided by one of the authors of the SON-R 5½-17. The individual administration of the subtests, which occurred at the school of the pupils, required approximately one hour.

The adaptive procedure of the SON-R was not used in this study. Instead, the items were administered in order of increasing difficulty. The administration of the subtests Categories and Situations was stopped after 12 errors; with the subtest Stories a stopping rule of eight errors was maintained. After each item the child was informed whether the answer was right or wrong. Providing feedback is an important part of the standard administration procedure of the SON-R, because it clarifies the instructions and gives the examinee the opportunity to learn from his errors and successes and to adjust his problem solving strategy. Immediately after an incorrect answer to an item, the children were asked whether they recognized and could name the pictures used in the item.

In Categories each item contains eight pictures: three example pictures that define the category and five alternatives from which to choose the two correct pictures. In the case of Situations, the subjects were asked if they recognized and could describe the main drawing and the missing parts. With Stories the children were asked to describe the pictures that had to be ordered. In addition to the administration of the three subtests, other data of the Brazilian children were gathered to obtain information about the validity of the test. For the 83 participants of the Brazilian sample, school marks on mathematics, science, and Portuguese were collected. The school teachers of the Brazilian children were requested to evaluate them on their degree of motivation, cooperation, and concentration during class hours. The evaluation was given on a 3-point scale, ranging from "low", *via* "average" to "high".

In the Netherlands, seven trained undergraduate psychology students administered Categories and Situations.

The authors of the SON-R 5½-17 provided supervision. The study in the Netherlands was realized to verify if problems of the Brazilian children with the recognition of determined pictures of Categories and Situations were really due to cultural bias or were caused by other factors. Examples of other causes of recognition problems are unclearly drawn pictures or pictures which represent objects that are infrequently used. Item bias was assumed to be present if one group showed more problems with the recognition of a determined picture than the other group. If both groups indicated considerable problems, this was an indication that the picture as such was difficult to recognize.

The subtest Stories was not administered in the Netherlands because the Brazilian children did not show any problems with the recognition of the pictures used in this subtest. In the Netherlands the same administration procedure as in Brazil was followed for the subtests Categories and Situations. The individual administration of the subtests, which occurred at the school of the pupils, required approximately three-quarters of an hour.

Data analysis

In all analyses age-corrected standard scores were used (*M*=100, *SD*=15). The standardized subtest scores were obtained using the computer program that is included with the SON-R 5½-17. In some analyses the difference between groups was described using d-ratios. The d-ratio expresses the difference between the means in units of the standard deviation of the samples. Coefficient lambda 2 of Guttman (λ_2) was chosen for the estimation of reliability because it does not underestimate reliability as much as coefficient alpha, especially in the case of short tests (Ten Berge & Zegers, 1978). Since the reliability coefficients λ_2 were calculated on base of samples that were heterogeneous in relation to age, a correction for the influence of age was applied. A second correction of the reliability coefficients was applied in relation to the variance of the standardized scores (Guilford & Fruchter, 1978). To test the significance of the difference in percentage of unknown pictures between the Brazilian and Dutch sample, the Fisher exact probability test was used (Siegel & Castellan, 1988).

In the analysis of Differential Item Functioning (DIF) the procedure of Bilog-MG was employed (Zimowski, Muraki, Mislevy & Bock, 1996). This procedure assumes that differential item functioning only extends to the difficulty of the items and not to the discriminating power. In other words, the assumption is made that the slope parameters (*a*-parameters) of the items are homogeneous across groups. The item difficulties are allowed to differ from one group to another. For the groups that are being compared different latent distributions are assumed. Bilog-MG estimates the DIF effects of the items as contrasts between the reference group and the so-called focal group(s). In our analysis the reference group was the standardization sample of the SON-R 5½-17 of 1,350 subjects from the Netherlands, and the focal group was the sample of 83 Brazilian subjects. An item was classified as a DIF item when the difference of the *b*-parameters in the reference and focal group was statistically significant at the 5% level.

Results

Overall performance

Means, standard deviations, reliability coefficients (λ_2) and d-ratios of the differences in mean scores are presented in Table 1. The Brazilian children obtained a lower mean score on the subtests Categories and Situations in comparison with the Dutch children. These differences are significant at the 5% level. According to Cohen"s classification (Cohen, 1992), the d-ratio of the difference between the mean score for the two groups for the subtest Categories indicates a small effect size, while the d-ratio for the subtest Situations suggests a medium effect size.

Table 1. Means, standard deviations and reliabilities (λ_2) on three SON-R subtests for the Brazilian and Dutch group and d-Ratios between the two groups

	Brazilian group (*N* = 83)			Dutch group (*N* = 51)			
	Mean	**(SD)**	λ_2	**Mean**	**(SD)**	λ_2	**d-Ratio**
Categories	94.8	(15.9)	.74	100.4	(16.5)	.75	-.35
Situations	95.0	(20.3)	.67	109.2	(15.2)	.71	-.77
Stories	97.5	(16.9)	.69	-	-	-	

Notes - The reliability coefficients λ_2 were corrected for age and for the standard deviations of the two groups.

- The d-ratio expresses the difference between the means of the Brazilian and Dutch group in units of the standard deviation (*SD*).

The higher d-ratio for Situations is a consequence of the relative high performance of the Dutch children on this subtest. Within the Brazilian group the differences between the three subtests are not statistically significant. The reliability coefficients λ_2 of .74 and .67 for Categories and Situations for the Brazilian children are quite similar to the values of .75 and .71 in the standardization sample of the Netherlands. The correlations between the subtests are relatively high in the Brazilian sample (Table 2). The correlations involving the subtest Situations are significantly higher at the 5% significance level for the Brazilian children than for the Dutch children. The higher correlations of the subtest Situations in the Brazilian sample might be related to the high standard deviation of this subtest (*SD* = 20.3) compared to the standard deviation of this subtest in the standardization sample of the Netherlands (*SD* = 15.0).

Table 2. Correlations (corrected for unreliability) between the three SON-R subtests for the Brazilian group and the Dutch standardization sample

	Brazilian group	**Dutch standardization sample**
	***N* = 83**	***N* = 1,350**
Categories - Situations	.77	.59
Categories - Stories	.58	.51
Situations - Stories	.84	.74

Note - With exception of the correlation between the subtests Stories and Categories, the correlations between the three subtests of the SON-R are significantly higher at the 5% level in the Brazilian sample than in the Dutch standardization sample.

Recognition of pictures

The first procedure to identify the presence of item bias was based on the recognition of the pictures by the children who gave a wrong answer to an item. The basic idea behind this procedure is that children should not fail an item because they are unfamiliar with one or more pictures used in that item. Table 3 displays the twelve items of the subtest Categories containing pictures unknown to at least 20% of the Brazilian or Dutch children who could not solve the item. Each item of the subtest Categories is composed of eight different pictures, three to define the category and five pictures from which two should be chosen that belong to the category. The second column of the table describes the pictures used in these items. The third column displays the number of Brazilian children who gave an incorrect answer to the item. The fourth column shows the percentage of these children that did not recognize the picture. The fifth and sixth column show the same information for the Dutch group. The last column shows the difference in percentage for the two groups and whether this difference is statistically significant at the 5% level.

According to the results of Table 3, pictures used in items 2b, 2c, 4b, 6a, and 8a were unknown to a higher percentage of the Brazilian group compared to the Dutch group. This is an indication that these five items are biased in favor of the Dutch group.

A higher percentage of the Dutch group encountered problems in the recognition of one of the pictures used in items 4a and 9a: these two items seem to be biased in favor of the Brazilian group. Both groups showed the same degree of problems recognizing pictures in items 1c, 3a, 3b, 5c, and 9c. These five items do not seem to be culturally biased since the pictures were difficult to recognize for both groups.

Table 4 presents the same type of information for the subtest Situations. Inspection of this table reveals that there are nine items of Situations with a picture unknown to at least 20% of the children who responded incorrectly to the item. The last column of the table shows that item 10a is the only

Table 3. Items of the subtest Categories which contain pictures unknown to at least 20% of the Brazilian or Dutch children who failed the item

Categories						
Item	**Picture**	**Brazilian group**		**Dutch group**		
		N-wrong	% unknown	N-wrong	% unknown	**Difference in %**
1c	A4 - (factory)	13	23.1	5	20.0	3.1
2b	A4 - (electric outlet)	20	50.0	8	0.0	50.0 *
2c	A3 - (bird nest)	26	34.6	16	0.0	34.6 *
3a	E1 - (stop watch)	30	26.7	13	38.5	-11.8
3a	E3 - (thermometer)	30	20.0	13	30.8	-10.8
3b	A1 - (dish rack)	29	6.9	18	33.3	-26.4
4a	E1 - (bolt of textile)	56	12.5	41	65.9	-53.4 *
4b	A2 - (wash cloth)	38	94.7	12	0.0	94.7 *
5c	A4 - (wash tub)	40	7.5	16	31.3	-23.8
6a	A1 - (sledge)	58	62.1	37	0.0	62.1 *
8a	E3 - (handlebars)	42	54.8	28	0.0	54.8 *
9a	A4 - (mosque)	47	8.3	12	33.3	-25.0 *
9c	A2 - (diagram)	18	50.0	25	44.0	6.0

Notes - E1, E2, and E3 are the examples that define the category; A1 to A5 are the alternatives to choose from.
- Positive differences indicate items relatively unknown to the Brazilian group.
- Differences significant at the 5% level are marked with an asterisk.

Table 4. Items of the subtest Situations which contain pictures unknown to at least 20% of the Brazilian or Dutch children who failed the item

Situations						
Item	**Picture**	**Brazilian group**		**Dutch group**		
		N-wrong	**% unknown**	**N-wrong**	**% unknown**	**Difference in %**
1c	A1 - (chimney)	14	14.3	4	25.0	-10.7
2b	D - (man with stick)	17	29.4	3	66.7	-37.3
3c	D - (bath)	29	31.0	2	38.5	-7.5
4a	A2 - (pincers)	26	23.1	7	0.0	23.1
4b	A4 - (vegetables)	22	28.2	7	14.3	13.9
9a	D - (angry mother)	29	6.9	20	20.0	-13.1
9b	D - (child sorts blocks)	40	5.0	23	21.7	-16.7
10a	D - (biking contest)	43	0.0	28	21.4	-21.4 *
10c	D - (construction site)	26	3.8	25	24.0	-20.2

Notes - D is the main drawing with one to four pieces missing; A1 to A4 are alternatives to choose from.
- Positive differences indicate items relatively unknown to the Brazilian group.
- Differences significant at the 5% level are marked with an asterisk.

item for which the difference was statistically significant at the 5% level. In other words, only one item of the subtest Situations is biased according to this procedure. This item shows bias in favor of the Brazilian group. The remaining eight items 1c, 2b, 3c, 4a, 4b, 9a, 9b, and 10c contained pictures that were difficult to recognize for both groups. For the subtest Stories no results are displayed as the Brazilian children did not report any problems with the recognition of pictures in this subtest.

Resuming, the results of this procedure indicate that of a total of 80 items that were investigated, eight items seem to be culturally biased: seven items of the subtest Categories and one item of the subtest Situations. Of these eight items, five items favored the Dutch group and three items favored the Brazilian group. Thirteen items contained pictures that were difficult to recognize for both groups.

Item difficulty

The second procedure to assess the presence of item bias was based on the difficulties of the items (*b*-parameters) according to Item Response Theory (IRT). For this analysis the procedure of Differential Item Functioning (DIF) of the software program Bilog-MG (Zimowski et al., 1996) was used. The reference group in this analysis was the Dutch standardization sample of the SON-R 5½-17 of 1,350 children, while the 83 Brazilian children formed the focal group.

The first step in this procedure was to evaluate for each of the three SON-R subtests which IRT-model fitted best the data of the reference group and the focal group combined. For the subtests Categories and Situations the three-parameter model with a fixed *c*-parameter ("guess" parameter) showed the best model fit, while for the subtest Stories the two-parameter model showed the best fit. The next step was to test whether a DIF model or a non-DIF model fitted the data best. In a DIF model the two groups are considered as two independent groups with different *b*-parameters, while in the non-DIF models the two groups are treated as one group.

Table 5. Model fit of different IRT-models for the three SON-R subtests

	Non-DIF model	DIF-model	Difference	D.F.	C.R.
	- 2 log likelihood	-2 log likelihood			
Categories	23,546	23,458	88	26	3.38*
Situations	26,760	26,600	160	32	5.00*
Stories	17,196	17,141	55	19	2.89*

Notes - The critical ratio (C.R.) is the ratio of the difference of the -2 log likelihood of the two models and the degrees of freedom (D.F.).

- When the critical ratio is greater than 1,96 it is statistically significant at the 5% level.

Table 5 displays the -2 log likelihood of the non-DIF model and of the DIF model. The values of the DIF model are lower, which indicates a better model fit. The statistical test of the model fit is based on the difference of the log likelihoods (Camilli & Shepard, 1994). The difference of the log likelihoods and its degrees of freedom are displayed in the last columns of Table 5. The ratio of this difference and the degrees of freedom is called the critical ratio (*C.R.*). When this ratio is greater than 1,96, it is statistically significant at the 5% level. Table 5 shows that for all three SON-R tests the DIF-model fits the data significantly better. Thus the two groups were considered to be independent, and the *b*-parameters were estimated separately for each group.

Table 6 displays the items with a significant difference in *b*-parameter for the two groups. Of a total of 80 investigated items, 19 items were identified as items with DIF. Of these items with DIF, eleven were in favor of the Brazilian group and eight in favor of the Dutch group. Five items of the subtest Categories showed DIF: two items in favor of the Dutch group and three items in favor of the Brazilian group. Of the subtest Situations, eleven items were identified as items with DIF: five items in favor of the Dutch children, and six items in favor of the Brazilian children. The three items of the subtest Stories that showed DIF were all easier for the Brazilian children.

Table 6. Items of the three SON-R subtests which show Differential Item Functioning (DIF) based on the item difficulties according to Item Response Theory

Categories			
Item	**Description item**	**Difference in *b*-parameter**	**Standard error**
2c	animals	-0.50 *	0.23
4c	toys	-0.42 *	0.20
6a	means of transport	-0.67 **	0.17
6b	fasteners	0.43 *	0.17
9c	signs	0.81 **	0.30
Situations			
Item	**Description item**	**Difference in *b*-parameter**	**Standard error**
1b	hunting a rabbit	-2.03 **	0.46
2b	playing with a dog	-0.97 *	0.47
2c	posting a letter	-1.46 **	0.35
3a	ruling traffic	0.83 *	0.35
3c	taking a bath	-2.02 **	0.44
4b	selling flowers	-0.59 **	0.18
5c	breaking dishes	0.45 **	0.17
7a	playing football	1.46 **	0.25
7b	watching the mirror	0.54 **	0.20
9c	jogging along the beach	0.65 **	0.18
10c	working in construction	0.97 **	0.32
Stories			
Item	**Description item**	**Difference in *b*-parameter**	**Standard error**
6a	getting water at the well	0.42 **	0.16
9a	relaxing at the beach	0.43 *	0.19
9b	rowing with a boat	0.42 *	0.17

Notes - Differences in *b*-parameters significant at the 5% level are marked with one asterisk; differences significant at the 1% level are marked with two asterisks.

- Positive differences in *b*-parameters refer to items that are more difficult for the Dutch children, while negative differences indicate items more difficult for the Brazilian group.

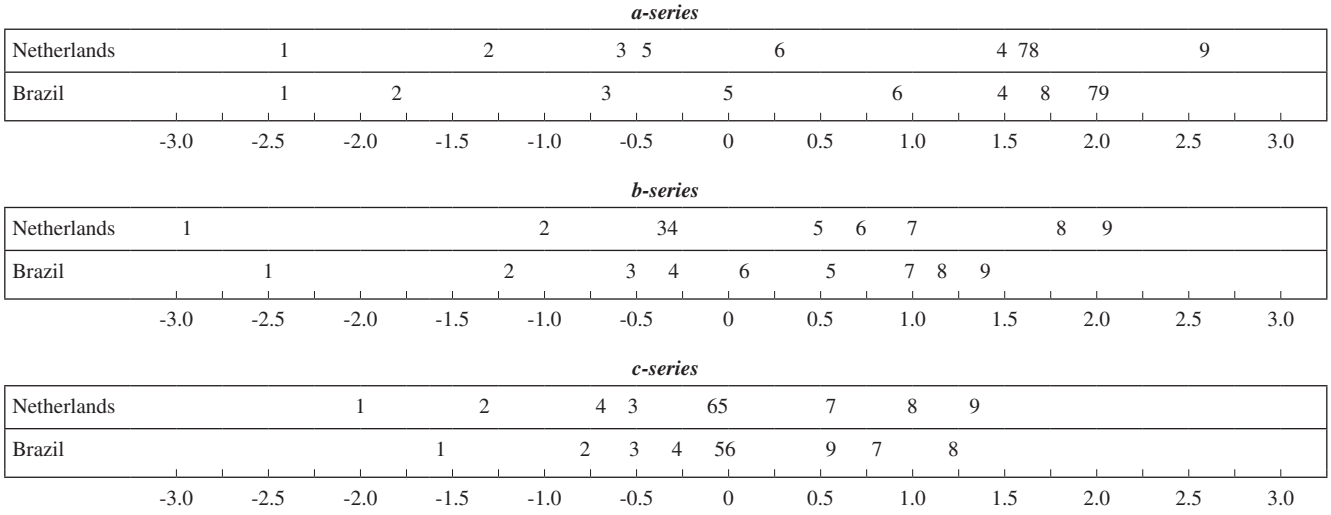

Figure 1. Plot of the *b*-parameters (item difficulties) of the items of subtest Categories for the Dutch standardization sample and the Brazilian focal group.

Correlations between indices of item difficulty

Despite the significant differences in *b*-parameters for specific items, there is a strong overall correspondence between item difficulties in the Brazilian group and the Dutch standardization sample. Table 7 shows that the correlation between the *p*-values of the two groups varies from .90 (Situations) to .98 (Categories). The correlation between the *b*-parameters is .87 for Situations and .96 for Stories and Categories. That the correlations between the *p*-values and between the *b*-parameters of the two groups give such similar results is not surprising. In the Dutch standardization sample the correlation between the *p*-value and the *b*-parameter is close to -.97 for all three subtests. As an example, Figure 1 shows in a visual way the strong correspondence between the *b*-parameters of the subtest Categories of the focal and the reference group. It also shows that item 4a is too difficult in both groups in relation to the order of administration and that for the Brazilian group item 9c is easier than items 7c and 8c.

Table 7. Correlations between different indices of item difficulty of the three SON-R subtests in the Brazilian group (N = 83) and in the Dutch standardization sample (N = 1,350)

	***p*-value Netherlands / *p*-value Brazil**	***b*-parameter Netherlands / *b*-parameter Brazil**	***p*-value Netherlands / *b*-parameter Netherlands**
	r	**r**	**r**
Categories	.98	.96	-.96
Situations	.90	.87	-.97
Stories	.95	.96	-.97

Note - All correlations are significantly different from zero at the 1% level of significance.

Validity

The Brazilian pupils were evaluated by their schoolteachers with respect to the degree of motivation, cooperation and concentration they display during school hours. The evaluation of the teachers was given on a 3-point scale, ranging from "low", via "average" to "high". The rationale behind this evaluation is that in the standardization research of the SON-R 5½-17 a moderate correlation of .33 was found to exist between the level of motivation, concentration and cooperation of the pupils at school and their performance on the SON-R intelligence test. An average correlation of .66 was found between the teacher's judgement and mean report marks. Apparently, factors such as concentration, motivation and cooperation are much more important for school achievement than for the performance on the SON-R intelligence test. Based on these findings, it was expected that also for the Brazilian pupils only moderate correlations would be found between the teacher's judgement on the degree of their motivation, concentration, and cooperation and their performance on the subtests of the SON-R 5½-17.

Table 8. Correlations of the three SON-R subtests with teacher's judgement of the motivation, concentration and cooperation of the Brazilian participants

	Motivation	**Concentration**	**Cooperation**
Categories	.39	.31	.38
Situations	.42	.31	.42
Stories	.32	.30	.25

Note - All correlations are significantly different from zero at the 1% level of significance.

Table 8 shows that the scores of the Brazilian children on the three subtests are, as expected, only moderately related to teacher's judgement of their degree of motivation, concentration, and cooperation. Situations showed the highest correlations with these characteristics and Stories the lowest. The moderate correlations for the Brazilian children are quite similar to the correlations found in the standardization research of the SON-R 5½-17 in the Netherlands (Snijders et al., 1989).

For a part of the Brazilian children and of the Dutch standardization sample, school marks on language and mathematics were available. Table 9 shows the correlations

Table 9. Correlations – corrected for unreliability – of the test scores on the three SON-R subtests with school marks on language and mathematics for a part of the Brazilian group (N=33) and for a part of the Dutch norm sample (N=490)

	Language		Mathematics	
	Brazilian group N = 33	Dutch group N = 490	Brazilian group N = 33	Dutch group N = 490
Categories	.44	.26	.60	.30
Situations	.32	.30	.41	.27
Stories	.33	.22	.41	.24

Notes - With the exception of the correlation between the subtest Categories and the school mark on mathematics none of the correlations differs significantly (at the 5% level) between the Brazilian and the Dutch group.

- All correlations are significantly different from zero at the 5% level of significance.

- corrected for unreliability - of the subtests with these school marks. In the Brazilian group, the correlations with school marks on language are slightly higher than in the Dutch group, although none of the differences is significant at the 5% significance level.

The correlation of .60 for the Brazilian group between the scores on the subtest Categories and the school marks on mathematics is significantly higher than the correlation of .30 for the Dutch group. Also for the other two subtests the correlations with mathematics are higher in the Brazilian group, although these differences with the Dutch group are not significant at the 5% level of significance.

Discussion

The results of the first procedure used in this study indicate that 21 of the 80 items of the subtests Categories, Situations and Stories contain pictures that are difficult to recognize for the Brazilian children, the Dutch children or for both groups. Thirteen of these problematic items are probably not culturally biased as both Brazilian and Dutch children reported problems recognizing these pictures. Possible explanations for the observed difficulties with 6 of these 13 items are: (a) use of old fashioned designs of the reproduced objects (stop watch, thermometer, dish rack); (b) inclusion of pictures representing old fashioned objects that are no longer in use (wash tub); or (c) inclusion of pictures that are simply hard to recognize (factory, diagram). For the other seven problematic items that were difficult to recognize for both groups no good explanations could be found.

There are clear indications that 8 of the 21 items are culturally biased. Five of these items are biased in favor of the Dutch group and three in favor of the Brazilian children. Various explanations can be given why a relatively great part of the Brazilian children did not recognize certain pictures. Obviously, some pictures were not recognized because the reproduced objects are uncommon in Brazil (washcloth, sledge), other pictures were not recognized because the design of the object is quite different in Brazil compared to the Netherlands (electric outlet, handlebars of a bicycle). A possible explanation why more Dutch than Brazilian children had difficulties recognizing the textile bolt might be that the shops in the Netherlands are more modern than the ones in Brazil and expose less frequently products like textile bolts. For the other two items with pictures that were difficult to recognize for the Dutch group no satisfactory explanation could be given.

With the procedure based on the IRT item difficulties, 19 items with DIF were identified: five items of Categories, 11 of Situations, and three of Stories. It is important to remark here that DIF indices as such do not provide immediate evidence of item bias. Content analysis of the items is required to judge the implications of DIF for cultural item bias. Especially in small samples, DIF statistics can produce incalculable Type I and Type II error rates (Camilli & Shepard, 1994). Therefore, after the DIF analyses we tried to find explanations for the differential functioning of items that could be associated with group membership.

Of the five items with DIF of the subtest Categories only for one item (item 6a) a convincing explanation could be given. This item showed the highest value of bias in favor of the Dutch children. The bias is most likely due to the inclusion of a sledge as one of the correct alternatives, an object that is seldom or never used in Brazil as a consequence of its climate. This item was also detected as biased with the first procedure. No good reasons could be found to explain why DIF occurred with the remaining four items.

Of the items of Situations that showed DIF in favor of the Dutch children, items 1b, 2c and 3c displayed relative large differences in item difficulties. In case of items 1b and 3c the bias might be explained by the fact that the displayed activities, hunting a rabbit and taking a bath in a bathtub are no regular activities in Brazil. In case of item 2c (posting a letter), the explanation lies in a different design of postboxes used in Brazil. For items 2b (playing with a dog) and 4b (selling flowers), the bias might be explained by the fact that the displayed activities are no regular activities in Brazil. Especially the poorer Brazilians do not usually keep dogs as pets, and flowers are seldom sold out in the open. Item 7a (playing football) showed a large difference in item difficulty. The bias in favor of the Brazilian children might be explained by the central role that football plays in Brazilian daily life. For the other items with bias in favor of the Brazilian children no convincing explanation for the occurrence of DIF bias could be found. For the three items with DIF of the subtest Stories no convincing explanations could be found.

Resuming, with the first procedure eight items were identified as biased, and with the second procedure seven items. Both procedures indicate item 6a as biased. Interestingly, the first procedure identified mainly items of Categories as biased, while the second procedure classified mainly items of Situations as biased. Altogether, 14 items were identified as biased. Of these, ten favored the Dutch children and four favored the Brazilian children. Taking into account that the total number of items investigated is 80, the negative effect of cultural bias for the Brazilian children is rather small.

One way to establish the effect of item bias is to analyze the correspondence in order of item difficulties between the Brazilian and Dutch children. The order of item difficulties is especially important for the SON-R 5½-17 since the subtests are administered in an adaptive way. For the effectiveness of the adaptive procedure, the order of item

difficulty is essential. Results of this analysis revealed that the correspondence of the item difficulty is rather high. The correlation between the classical item difficulty (*p*-value) in Brazil and the Netherlands varies from .90 to .98. The correlation between the item difficulty based on IRT varies between .87 and .96. The weakest correspondence in order of item difficulty between the two compared groups was found for the subtest Situations. Apparently, the effect of item bias on the order of item difficulty in this subtest was stronger than in the other two subtests.

A basic question to be answered in this study was to which extent the occurrence of item bias influenced the validity of the test. Or, in other words: what are the practical consequences of item bias for the valid use of the SON-R test? The results show that the validity of the three subtests in the Brazilian group is highly comparable to the validity in the Netherlands. Correlations of the subtest scores with teacher's judgement of the motivation, cooperation and concentration of the Brazilian children are quite similar to the correlations with these characteristics found in the Netherlands (Snijders et al.,1989). The correlations of the subtest scores with school marks on language and mathematics in Brazil are also quite similar to those found in the Netherlands, with exception of the correlation found between the subtest Categories and mathematics which is significantly higher for the Brazilian sample at the 5% level of significance. All three subtests of the SON-R show a high predictive validity of school performance in the Brazilian sample. Of course, it should be remarked here that the present study uses only a small sample of Brazilian children, and that for full validity evidence for the SON-R 5½-17 more validity studies with larger samples sizes should be conducted. Studies with larger samples sizes and all subtests of the SON-R 5½-17 would also allow the verification of the equality of its factor structure across Dutch and Brazilian samples. These studies would provide more evidence for the cross-cultural validity of the SON-R 5½-17.

Although the results of the present study indicate that the test can be used in Brazil in its current form, it provides valuable information on how to improve the pictorial contents in the next revision. Even if these changes might hardly effect the psychometric qualities of the test, the face validity will improve in so far as the contents become less dominated by Western European life style. The present study also made apparent that some items of the subtests Categories and Situations have become outdated for use in the Netherlands and in Brazil. These results will be incorporated in the next revision of the SON-R 5½-17 as well.

References

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

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin.

Cohen, J. (1992). Quantative methods in psychology: A power primer. *Psychological Bulletin, 112*(1)*,* 155-159.

Guilford, J. P. & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). New York: McGraw Hill.

Helms-Lorenz, M. & Van de Vijver, F. J. (1995). Cognitive assessment in education in a multicultural society. *European Journal of Psychological Assessment, 11*(3)*,* 158-169.

Hu, S. & Oakland, T. (1991). Global and regional perspectives on testing children and youth: an empirical study. *International Journal of Psychology, 26*(4)*,* 329-344.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Laros, J. A. & Tellegen, P. J. (1991). *Construction and validation of the SON-R 5½-17, the Snijders-Oomen non-verbal intelligence test*. Groningen: Wolters-Noordhoff.

Muñiz, J., Prieto, G., Almeida, L. & Bartram, D. (1999). Test use in Spain, Portugal and Latin American Countries. *European Journal of Psychological Assessment, 15*(2), 151-157.

Oakland, T., Wechsler, S., Bensuan, E. & Stafford, M. (1994). The construct of intelligence among Brazilian children – An exploratory study. *School Psychology International, 15*(4)*,* 361-370.

Siegel, S. & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). Tokyo: McGraw-Hill.

Snijders, J. T., Tellegen, P. J. & Laros, J. A. (1989). *Snijders-Oomen Nonverbal Intelligence Test, SON-R 5½-17, Manual & Research Report.* Lisse: Swets & Zeitlinger.

Tellegen, P. J. & Laros, J. A. (1993). The construction and validation of a nonverbal test of intelligence: the revision of the Snijders-Oomen tests. *European Journal of Psychological Assessment, 9*(2)*,* 147-157.

Tellegen, P. J., Winkel, M., Wijnberg-Williams, B. J. & Laros, J. A. (1998). *Snijders-Oomen Nonverbal Intelligence Test, SON -R 2½-7, Manual & Research Report.* Lisse: Swets & Zeitlinger.

Ten Berge, J. M. F. & Zegers, F. E. (1978). A series of lower bounds to the reliability. *Psychometrika, 43*(4)*,* 575-579.

Van de Vijver, F. J. R. & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*(2)*,* 89-99.

Van de Vijver, F. J. R. & Poortinga, Y. H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? *European Journal of Psychological Assessment, 8*(1)*,* 17-24.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4)*,* 473-492.

Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R. D. (1996). *Bilog-MG: Multiple group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.

Recebido em 14.06.2004
Primeira decisão editorial em 06.07.2004
Versão final em 14.07.2004
Aceito em 23.07.2004 ■