A method is presented for evaluating the presence and size of cross-cultural item biases. The examined items concern parental support and family cohesion in a Likert-type questionnaire for adolescents in The Netherlands. Each evaluated item has two versions, a collectivist and an individualistic one, that measure the same theoretical construct. The standardized difference between the score means of the item versions, called the Δe score, gives an indication of the cultural bias of the item. As expected, most items were found to yield a higher Δe when respondents scored low on an individualistic scale for acculturation or originated from countries that are (more) collectivist. This procedure is recommended for use in testing items in pilot studies.

ASSESSING CROSS-CULTURAL ITEM BIAS IN QUESTIONNAIRES
Acculturation and the Measurement of Social Support and Family Cohesion for Adolescents

DIANNE A. VAN HEMERT
CHRIS BAERVELDT
MARJOLIJN VERMANDE
Utrecht University

The importance of cross-cultural validation can be illustrated by an incident with the Self-Reporting Questionnaire (SRQ), a psychiatric case-finding instrument developed by the World Health Organization to detect psychiatric patients among visitors to medical clinics. The questionnaire is designed especially for developing countries. When Kortmann (1990) examined the content validity of the answers to the questions of the SRQ in Ethiopia, he found a number of striking results. For example, he concluded " 'Do you feel unhappy?', a basic question in the diagnosis of depression, was associated for many Ethiopians with feelings of mourning from the loss of someone or someone's dying. This became evident as witnessed by the often-heard, spontaneous comment accompanying a 'no' answer on this question: 'No, because no one has died.' The concept 'unhappy' does not appear to exist in the Ethiopian culture unless there is a clear cause for it" (p. 386).

The SRQ underlines the fact that when questionnaires are used for respondents with a different cultural background, it is necessary to study their cross-cultural validity. Before researchers can make any statement about the social and psychological processes they want to study, they need to be sure that no bias is threatening the validity of their measurements. This conclusion has led to many studies on item bias. In cross-cultural psychology, item bias refers to every difference in an observed score for which no corresponding difference can be found in the psychological domain to which the scores are generalized (Poortinga & Malpass, 1986; also see Van de Vijver & Poortinga, 1997). Van de Vijver and Poortinga (1997) identified the most common causes of item bias. These included poor item translations and inadequate item formulations (e.g., complex wording). Also, contents of items are

AUTHORS' NOTE: We would like to thank Herbert Hoytink, Quinten Raaijmakers, Ronan Van Rossem, John Adamopoulos, and two anonymous reviewers of this journal for their extensive and useful comments. Correspondence should be directed to Chris Baerveldt, Faculty of Social Sciences, Utrecht University, PO Box 80140, 3508 TC Utrecht, The Netherlands, phone +31-30-2534687; e-mail: C.Baerveldt@fss.uu.nl.

JOURNAL OF CROSS-CULTURAL PSYCHOLOGY, Vol. 32 No. 4, July 2001 381-396

not always appropriate in all cultures. Usually, studies examining item bias (e.g., Holland & Wainer, 1993; Tanzer, 1991) concern tests and are of a psychometric nature (see Poortinga, 1995).

However, in this study the focus is on another kind of bias, namely, on the effects of cultural background on the mean scores of scales or items. The starting point will be an interpretation problem that researchers encounter who study cultural differences through a survey. What does it mean that adolescents of different cultural backgrounds respond differently to an item such as "My father supports me when I have problems"? A mean difference between cultural groups could be caused by the fact that adolescents of one cultural group really experience more support than the adolescents of the other, but may also be explained by differences in measurement errors possibly caused by social desirability or interpretation problems. In this article, a cross-cultural bias is at hand when the difference between mean scores on an item of different cultural groups is not (completely) caused by differences in the social reality of the members of these groups but (also) by differences in response behavior. The main question examined in this article is how to study this kind of bias.

A method will be presented to study item bias in surveys. In this method, items are biased on purpose, either in the main survey study or in a pilot study. The evaluation of this exercise can be used to design the final items for the follow-up study. The examined items are formulated in two versions (I and II) in which each version reflects a different cultural background (A and B). Under certain conditions, which are explained below, the standardized differences between the scores of these versions can be interpreted as a statistic that indicates the size of the bias. To illustrate the above method for assessing cross-cultural item bias, a survey study among 1,317 adolescents from Dutch high schools will be presented.

TESTING CROSS-CULTURAL BIASES: THEORY

For testing cross-cultural bias, it is useful to represent the measurement of a theoretical construct, X, through an item as follows:

$$Y = X + \varepsilon \tag{1}$$

where Y represents the score on the item, X is the 'true value' of the construct, and ε is the measurement error. A cross-cultural bias of an item exists when the measurement error of that item differs for subpopulations A and B with different cultural backgrounds:

$$\varepsilon(A) \neq \varepsilon(B) \tag{2}$$

The starting point is the idea of two versions of items being designed in a pilot study. The two versions of the item have to reflect two different cultural backgrounds. This requires first the study of the (differences between) cultural backgrounds through qualitative data and by comments of key persons of the studied groups on drafts of the items. Version I represents cultural background A, and version II represents cultural background B. Versions I and II are both numerical items within the same categories. Version I is tested on n_1 respondents, version II on n_2 respondents, randomly drawn from the total population (see Table 1).

Now, Y_1 represents the score of item version I on construct X_1 with measurement error ε_1, and Y_2 represents the score of item version II on construct X_2 with measurement error ε_2. Concluding, two versions of the same item are designed to represent the same construct X. Under this condition, we consider, following Van de Vijver and Leung (1997, p.8), the two versions

TABLE 1
Strategy for a Pilot Study: Two Item Versions in Two Subpopulations

Item Version	*Subpopulation A*	*Subpopulation B*	*Total Population*
Version I	$n_1(A)$	$n_1(B)$	n_1
Version II	$n_2(A)$	$n_2(B)$	n_2
Total	n(A)	n(B)	N

TABLE 2
Statistics for Two Item Versions and Their Differences

Item Version	*Subpopulation A*	*Subpopulation B*	*Difference*
Version I	$m_{Y1}(A) = mX1(A) + m_\varepsilon 1(A)$	$m_{Y1}(B) = mX1(B) + m_\varepsilon 1(B)$	$m_{Y1}(A) - m_{Y1}(B)$
Version II	$m_{Y2}(A) = mX2(A) + m_\varepsilon 2(A)$	$m_{Y2}(B) = mX2(B) + m_\varepsilon 2(B)$	$m_{Y2}(A) - m_{Y2}(B)$
Difference	$\Delta m(A) = mX1(A) - mX2(A) + m_\varepsilon 1(A) - m_\varepsilon 2(A)$	$\Delta m(B) = mX1(B) - mX2(B) + m_\varepsilon 1(B) - m_\varepsilon 2(B)$	$\Delta m(A) - \Delta m(B)$
Difference when $X_1 = X_2$ (structural equivalence)	$\Delta m(A) = m_\varepsilon 1(A) - m_\varepsilon 2(A)$	$\Delta m(B) = m_\varepsilon 1(B) - m_\varepsilon 2(B)$	$\Delta m(A) - \Delta m(B) = [m_\varepsilon 1(A) - m_\varepsilon 1(B)] - [m_\varepsilon 2(A) - m_\varepsilon 2(B)]$
Item, standardized	$\Delta e(A) = \Delta m(A)$/weighted s_i's	$\Delta e(B) = \Delta m(A)$/weighted s_i's	$\Delta e(A) - \Delta e(B)$

as being structurally equivalent to the other. Formally, every pair of items is called structurally equivalent when the following condition is met:

$$X_1 = X_2 \tag{3}$$

Consider the mean scores $mYi(A)$, with i = 1,2, representing item versions I or II for subpopulation A, and $mYi(B)$, with i = 1,2, representing item versions I or II for subpopulation B (see Table 2). The measurement errors are represented by $m_\varepsilon 1(A)$ for subpopulation A and by $m_\varepsilon 2(B)$ for subpopulation B. A possible cross-cultural bias can be evaluated by examining the differences between the mean scores for versions I and II. The difference $\Delta m(A)$ of the mean scores of versions I and II for subpopulation A is represented by

$$\begin{aligned}\Delta m(A) &= mY1(A) - mY2(A) = [mX1(A) + m_\varepsilon 1(A)] - [mX2(A) + m_\varepsilon 2(A)] \\ &= mX1(A) - mX2(A) + m_\varepsilon 1(A) - m_\varepsilon 2(A)\end{aligned} \tag{4a}$$

and analogously, the difference Δm_B for subpopulation B is represented by

$$\Delta m(B) = mX1(B) - mX2(B) + m_\varepsilon 1(B) - m_\varepsilon 2(B). \tag{4b}$$

Equation 4a implies that a difference score Δm_A that is not unequal to 0 does not necessarily point to a cross-cultural bias. Theoretically, the Δm_A score could be caused by real differences between the constructs underlying the versions I and II as well as by cross-cultural bias. However, because of the premise that both versions refer to the same construct $X_1 = X_2$, (see also Table 2), the expression (4a) changes into

$$\begin{aligned}\Delta m(A) &= mX1(A) - mX2(A) + m_\varepsilon 1(A) - m_\varepsilon 2(A)\\ &= 0 + m_\varepsilon 1(A) - m_\varepsilon 2(A)\\ &= m_\varepsilon 1(A) - m_\varepsilon 2(A),\end{aligned} \tag{5a}$$

and for equivalent reasons,

$$\Delta m(B) = m_\varepsilon 1(B) - m_\varepsilon 2(B). \tag{5b}$$

It is crucial for the reasoning in equation 4, and for the whole procedure, that $X_1 = X_2$, that I and II are items that measure the same theoretical construct, which are structurally equivalent. It is therefore necessary that this structural equivalence of versions I and II be confirmed empirically (see the next sections for more details).

Consider now the difference between $\Delta m(A)$ and $\Delta m(B)$ for the subpopulations A and B, respectively:

$$\begin{aligned}\Delta m(A) - \Delta m(B) &= [m_\varepsilon 1(A) - m_\varepsilon 2(A)] - [m_\varepsilon 1(B) - m_\varepsilon 2(B)]\\ &= m_\varepsilon 1(A) - m_\varepsilon 2(A) - m_\varepsilon 1(B) + m_\varepsilon 2(B)\\ &= [m_\varepsilon 1(A) - m_\varepsilon 1(B)] - [m_\varepsilon 2(A) - m_\varepsilon 2(B)]\end{aligned} \tag{6}$$

When $\Delta m(A) \neq \Delta m(B)$, it can be concluded from equation 5 that $m_\varepsilon 1(A) - m_\varepsilon 1(B) \neq 0$ and/or that $m_\varepsilon 2(B) - m_\varepsilon 2(A) \neq 0$. According to equation 2, this means that at least one of the versions, I or II, is cross-culturally biased. This has practical value because it means that the items should be changed when the absolute difference between the $\Delta m(A)$ and $\Delta m(B)$ is substantial. However, conversely one cannot be sure that there is no cross-cultural bias when $\Delta m(A) = \Delta m(B)$. Consider, for instance, the case that versions I and II have equal biases. According to equation 2, this can be expressed as $m_\varepsilon 1(A) - m_\varepsilon 1(B) = m_\varepsilon 2(A) - m_\varepsilon 2(B)$. According to equation 6, $\Delta m(A) - \Delta m(B) = [m_\varepsilon 1(A) - m_\varepsilon 1(B)] - [m_\varepsilon 2(A) - m_\varepsilon 2(B)] = 0$. In other words, a difference between $\Delta m(A)$ and $\Delta m(B)$ is an indication of cross-cultural bias, but the absence of a difference between $\Delta m(A)$ and $\Delta m(B)$ is no guarantee of cross-cultural validity. Therefore, it seems that we can never be sure that there is no cross-cultural bias.

It is possible to overcome this problem by using a strategy of intentionally designing cross-cultural biases in pilot studies. In this study, this was done by designing version I and II in such a way that version I focuses more on a description of the cultural background of subpopulation A, whereas version II focuses more on a description of the cultural background of subpopulation B. This requires a theoretical basis, of which an elaborate example is presented in the next section. Consider the simple example of a construct *X* of empathy, measured by a population comprising respondents with a visual handicap (A) and one comprising respondents without a visual handicap (B). Version I is formulated as "I can feel when people are lying," version II as "I can see when people are lying." The items are formulated positively (the respondents score higher on items when they agree more with the item) and are symmetric (the negative categories are exact negations of the positive categories). When using this strategy, it can be demonstrated that the size of the difference between $\Delta m(A)$ and $\Delta m(B)$ is an indication of the size of the cross-cultural bias.[1] For the interpretation of the difference between $\Delta m(A)$ and $\Delta m(B)$ it is useful to standardize both statistics by the weighted standard deviations of the item versions. Because of its relation to error terms, this standardized statistic is named Δe (see Table 2). For population A this is expressed by

$$\Delta e(A) = \frac{(m_{Y1}(A) - m_{Y2}(A))}{\sqrt{\left(\frac{((n_1(A)-1)S_1(A)^2 + (n_2(A)-1)s_2)}{n_1(A)+n_2(A)-2)}\right)}} \quad (7a)$$

and for population B by

$$\Delta e(B) = \frac{(m_{Y1}(B) - m_{Y2}(B))}{\sqrt{\left(\frac{((n_1(B)-1)s_1(B)^2 + (n_2(B)-1)s_2(B)^2)}{(n_1(B)+n_2(B)-2)}\right)}} \quad (7b)$$

where m$Y1$ is the mean score of version I and $mY2$ the mean score of version II. The s_i stands for the standard deviations of versions I and II and the n_i for the numbers of cases on which the versions are tested. Δe has $n_1 + n_2 - 2$ degrees of freedom, and its statistical significance can be evaluated (for example, see Ott, Mendenhall & Larson, 1978, pp. 257-271). Moreover, Δe resembles Cohen's *d*, giving a basis to evaluate its size (Cohen, 1992). However, the size of Δe is not important, but the size of the difference between $\Delta e(A)$ and $\Delta e(B)$ is. Therefore, following Cohen, $\Delta e(A) - \Delta e(B)$ is considered small when it is about .20, medium when it is about .50, and large when it is greater than .80. It is possible to test the significance of differences between Δe scores by analysis of variance.[2] When Δe(A) – $\Delta e(B)$ is small for items in a pilot study, versions I and II can be used in the final study. When $\Delta e(A) - \Delta e(B)$ is large and significant, the researcher should formulate new items.

APPLICATION

It is crucial for the above procedure that $X_1 = X_2$, that is, that versions I and II are structurally equivalent. This may cause some confusion. Consider, for instance, the following two items: (a) "My father helps me with my homework" and (b) "My father helps me repair my bike". It could be argued that (a) and (b) both reflect the construct of practical support from father. However, it could also be argued that the first item reflects the father's interest in school performance and the second item does not. This example illustrates that the criterion that two items be structurally equivalent depends on the construct or scale to which the items are supposed to relate. Therefore, it makes sense to study the structural equivalence of versions of items to a specific scale, that is, to study whether both versions could be a comparable item in that scale. For one-dimensional scales this comes down to item-scale correlations of the same size. When scales are multidimensional, the correlations with all factors or factor loadings should be compared.

When structural equivalence has been made probable, the procedure of analysis seems easy. Respondents can be split up according to ethnicity, nationality, or native country, and the Δe scores could be compared between those groups. However, such a comparison would oversimplify the question of culture (Hofstede, 1991). Berry, Trimble, & Olmedo (1986) illustrated this point by stating that contradictory results can be found when studies are replicated: Different samples from the same cultural population may still differ. Therefore, it seems that a better procedure is to compare respondents on a major cultural dimension. Many researchers (e.g., Hofstede, 1980; Triandis, 1990) discriminate between individualistic

cultures and collectivist ones. There is some discussion about this discrimination, because individualistic and collectivist features are not logical opposites of each other. However, for the purpose of this demonstration and for reasons of expediency, we will assume that these two kinds of cultures can be imagined as two extremes on one dimension. By definition, a person in an individualistic culture is mainly motivated by personal choices, goals, and achievements, whereas someone in a collectivist culture is more subject to group rules. For example, according to Triandis et al. (1993), independence and creativity are emphasized in individualistic cultures, whereas obedience and cooperation are emphasized in collectivist cultures.

The distinction between collectivist and individualist cultures can be represented by values because it offers a basis to compare, for example, immigrants and indigenous people. However, for immigrants their home culture is just a starting position. Immigrants face a process of acculturation, that is, a process that changes individuals, either through contacts with a different culture or as a result of the changes experienced by their own culture because of acculturation (Berry, 1990). The level of acculturation of immigrants may vary independently of their nationality or ethnic group. Therefore, it is useful to test cross-cultural validity by not merely comparing Δe scores between ethnic group or nationality but also between individual levels of acculturation. In general, Berry et al. (1986) state that the measurement of acculturation adds to the validity and reliability of research.

HYPOTHESES

In this study, the method described above was tested on a survey of adolescents in Dutch, urban, secondary schools. There is a great variety in the adolescents' levels of acculturation and ethnic backgrounds in urban schools in The Netherlands. Most immigrant children come from three main immigrant groups: Morocco, Turkey, and Surinam. The cultural background of adolescents with parents from Morocco or Turkey can mainly be regarded as collectivist. Their parents came to The Netherlands in the 1960s when the Dutch economy required additional labor forces and therefore employed unskilled workers. The Surinamese culture is regarded as halfway along the continuum of individualism and collectivism. The Surinamese were the first large group to immigrate after the Second World War. As inhabitants of a former Dutch colony, they spoke the Dutch language. The Surinamese mostly came to The Netherlands for a good education (Van Niekerk, 1993). Dutch culture is mainly regarded as individualistic. The individualistic-collectivist dimension is reflected in the way respondents view their personal relations and personal network. Therefore, it can be expected that respondents from the various ethnic groups will respond differently on items about parental support and family cohesion. When these items are formulated in such a way that they are more a reflection of the dominant Dutch culture, it is to be expected that the items will have less item bias when the respondents have become more acculturated. Surinamese people, with a history of colonization by the Dutch, are more familiar with the Dutch culture and education system than Turkish and Moroccan people. As such, Surinamese adolescents are expected to be more acculturated than Turkish and Moroccan adolescents but less than Dutch adolescents. In several Dutch studies, Surinamese people were found to take a middle position between Dutch people and both Turkish and Moroccan with respect to their educational goals (Janssens, Pels, Dekoviç, & Nijsten, 1999; Pels, 1994). Consequently, adolescents with Moroccan or Turkish roots are expected to have the most extreme biases, followed by Surinamese and Dutch adolescents, respectively.

METHOD

RESPONDENTS

The participants were composed of 1,317 pupils of 20 urban high schools (aged 16-18 years, sexes equally represented). The respondents completed a questionnaire. High school pupils in The Netherlands can decide between several levels of secondary education. Schools often cover several levels, but the classes in school mainly consist of pupils from the same level. The respondents all studied at an intermediate level of secondary education, the so-called MAVO. At the MAVO level pupils study languages and sciences and they also learn some basic technical subjects.

The cultural backgrounds of the pupils differed: 64.5% of all respondents had two parents who were born in The Netherlands, 4.7% had Moroccan origins, 5.7% had Turkish, and 7.6% had Surinamese parents. Ethnicity was measured as the country of birth of both parents. For example, a respondent was considered to be Turkish when both of his or her parents came from Turkey. Children from parents with other than Dutch, Surinamese, Turkish, or Moroccan origins (6.7%) and children from mixed marriages (10.7%) were not included in this part of the study. The parents of the Dutch pupils were mainly nonreligious (58.6%) or Christian (35.3%), whereas 98.4% of the Moroccan and 91.9% of the Turkish pupils had Islamic parents. Most (52.6%) Surinamese pupils came from Hindu families and the rest were Christian, Islamic, or nonreligious. The Dutch respondents lived in smaller families than the Turkish and Moroccan pupils. Half of the Turkish parents were born in nonurban areas, whereas two thirds of the parents of other ethnic groups grew up in urban areas.

MEASURES

The examined items were related to three scales (Social Support by Father, Social Support by Mother, and Family Cohesion). Acculturation was measured by a scale of four items regarding the use of the Dutch language and Dutch media at home. Cronbach's alpha for this scale was .74. The scale was one dimensional: A factor analysis resulted in one factor that explained 59.4% of the variance and had an eigenvalue of 2.38. However, the scale was distributed in an extremely skewed way. Most respondents scored (almost) maximally on all items of the scale because they were strongly acculturated. A minority scored extremely low, indicating a serious distance from the dominant culture, whereas a group of pupils could also be recognized as partially acculturated. For these reasons the respondents were split up according to three categories: low acculturation ($n = 114$), medium acculturation ($n = 371$) and high acculturation ($n = 767$).

Three scales were used for studying structural equivalence: Support by Father, Support by Mother, and Family Cohesion. The first scale consisted of seven items. It had sufficient internal cohesion (Cronbach's alpha = .84) and could be considered one dimensional (eigenvalue = 3.64, explaining 52.0% of the variance). The Support by Mother scale (seven items) was equally reliable, with a Cronbach's alpha of .82 and a one-factor solution with eigenvalue 3.43 (49.0% explained variance). The Family Cohesion scale consisted of seven items. It had a sufficient .78 Cronbach's alpha. This scale could also be considered one dimensional (eigenvalues 3.07, 1.25; the first factor explains 43.9%).

PROCEDURE

The research sample was split up by randomly mixing the questionnaire in two versions. In version I, 8 support items and 8 cohesion items were designed to be more individualistic; in version II, these 16 items were designed to be more collectivist. In version I, the individualistic version, the items were designed to be problematic to pupils with a low degree of acculturation and a strong collectivist cultural background. In version II, the collectivist version, each item corresponded with an item of version I: The items were constructed to be structurally equivalent. However, in version II the items were formulated in such a way that respondents of lower acculturation levels should also be able to answer them without considerable effort. The items were also designed to correlate with either one of the social support scales or with the Family Cohesion scale consisting of items that were not manipulated.

Note that the manipulated items did not belong to the support scales and the Family Cohesion scale. The scales were not manipulated in any way and were the same in both versions. The manipulated items were formulated in such a way that they could be items of the scales. Whether these items correlated sufficiently with the scales had to be tested.

In formulating the collectivist items, the following instructions of Brislin (1986) were consulted for writing items for cross-cultural research. First, a general prescription is that items should not contain double negations, passive voices, or long sentences. Second, they should neither contain sayings, expressions, nor proverbs. In addition, the content of the collectivist items was verified and asserted by a pilot study in which respondents were asked for comment. The comments of a number of field workers were also used. The individualistic items were designed to be the opposite of the collectivist items, that is, Dutch sayings, expressions, and proverbs were used on purpose and in general they were meant to reflect typical features of the dominant Dutch culture. Table 3 shows all items with their versions I and II.

The reasoning behind some of these items will shortly be discussed. For instance, the versions of Item FC5 differed with respect to the reason for celebration. Knowing that birthdays are hardly celebrated in Islamic countries such as Turkey and Morocco, but that family gatherings are nonetheless important, the two versions of FC5 can be considered as measuring the same construct (i.e., celebrating with the family) while at the same time a culture-specific dimension is added. FC7 is an example of a very subtle difference in meaning, nevertheless conveying considerable cultural loading. In individualist cultures people are primarily taught to be independent, to be able to look after themselves. In collectivist cultures, however, people are mainly focused on their environment and are expected to be interdependent.

RESULTS

Acculturation was distributed as expected. The Dutch were significantly ($F = 484.7$, $df = 1032$, $p < .001$) more acculturated than the other groups, followed by the Surinamese pupils, whereas the Turkish and Moroccan groups came last. There were no significant differences between boys and girls here.

THE STRUCTURAL EQUIVALENCE OF VERSIONS I AND II

Table 4 shows the correlations of the two versions of the manipulated items with the (nonmanipulated) scales. Most correlations were positive, as expected. However, versions I

TABLE 3

Individualistic and Collectivist Versions of All Manipulated Items

Code	*Content*	*Individualistic Version*	*Collectivist Version*
SF1	Outdoor activities	I do outdoor activities with my father, like going to the movies or eating out in town.	I do outdoor activities with my father, like going to the market or visiting family.
SF2	Parental advice	My father advises me about contraceptives.	My father advises me about what I should do after school.
SF3	Organizing a party	My father and I organize birthday parties together.	My father and I organize family parties together.
SF4	Supporting	When my father thinks something is wrong at school, he will discuss this with the teacher.	When there is gossiping going on about our family, my father tries to do something about it.
SM1	Outdoor activities	I do outdoor activities with my mother, like going to the movies or eating out in town.	I do outdoor activities with my mother, like going to the market or visiting family.
SM2	Parental advice	My mother advises me about contraceptives.	My mother advises me about school and career.
SM3	Organizing a party	My mother and I organize birthday parties together.	My mother and I organize family parties together.
SM4	Supporting	When my mother thinks something is wrong at school, she will discuss it with the teacher.	When there is gossiping going on about our family, my mother tries to do something about it.
FC1	Having to take care of oneself	In our family everyone is responsible for themselves.	In our family everyone must take care of themselves.
FC2	Contact with other families	We often have coffee with the neighbors.	We often drop by at other families in the neighborhood.
FC3	Rules within the family	It is important that we all stick to the arrangements in our family.	It is important that we all stick to the rules in our family.
FC4	Amount of conversation	At home we talk a lot about issues such as the environment, unemployment, and refugees.	At home we talk a lot about our friends and family.
FC5	Festivities	When it is someone's birthday, we celebrate with the whole family.	When there is something to celebrate, the whole family participates.
FC6	Making decisions	My parents help me to make important decisions.	My parents make important decisions for me.
FC7	Trust	In our family we can rely on each other.	In our family we have to rely on each other.
FC8	Relations between family members	The family members are committed to each other.	The family members depend on each other.

TABLE 4
Correlations of Individualistic and Collectivist Versions of All Manipulated Items with Corresponding Scale

			Individualistic Version		*Collectivist Version*		*Difference Between Correlations*	
Scale		*Item*	n	*Correlation, Item With Scale*	n	*Correlation, Item With Scale*	t	p *(2-tailed)*
Father's support	SF1	Outdoor activities	473	.64**	615	.62**	0.32	.75
Father's support	SF2	Parental advice	473	.50**	616	.62**	–2.77	<.01
Father's support	SF3	Organizing a party	215	.51**	355	.44**	1.08	.28
Father's support	SF4	Supporting	258	.47**	239	.35**	1.52	.13
Mother's support	SM1	Outdoor activities	478	.65**	634	.56**	2.39	.02
Mother's support	SM2	Parental advice	474	.46**	634	.56**	–2.31	.02
Mother's support	SM3	Organizing a party	218	.45**	371	.36**	1.31	.19
Mother's support	SM4	Supporting	259	.46**	245	.29**	2.18	.03
Family cohesion	FC1	Having to take care of oneself	483	.43**	633	.36**	1.44	.15
Family cohesion	FC2	Contact with other families	480	.14	547	.23**	–1.59	0.11
Family cohesion	FC3	Rules within the family	480	.03	626	.23**	–3.22	<.01
Family cohesion	FC4	Amount of conversation	478	.09	633	.19*	–1.68	.09
Family cohesion	FC5	Festivities	476	.19*	631	.20*	–0.10	0.92
Family cohesion	FC6	Making decisions	478	.34**	630	.17*	2.93	<.01
Family cohesion	FC7	Trust	479	.27**	627	.26**	0.11	0.91
Family cohesion	FC8	Relations between family members	481	.33**	628	.28**	0.98	.33

NOTE: 2-tailed significance of correlations: *p <.01; **p <.001.

TABLE 5
Statistics for Two Versions of Item FC5 From the Family Cohesion Scale

	Subpopulation by Degree of Acculturation			*Subpopulation by Ethnic Group*			
	High	*Medium*	*Low*	*Dutch*	*Surinamese*	*Turkish*	*Moroccan*
Individualistic version							
Score	4.60	4.25	3.29	4.51	4.40	3.35	4.09
Number	274	134	45	298	42	26	23
Collectivist version							
Score	3.83	3.80	3.76	3.78	4.30	3.85	4.09
Number	377	169	59	392	47	40	32
Difference	0.77	0.45	–0.47	0.73	0.10	–0.5	–0.01
Standardized difference, Δe	0.78	0.46	–0.48	0.75	0.11	–0.51	–0.01

of the items FC2, FC3, and FC4 were not significantly correlated to the Family Cohesion scale. Therefore, the structural equivalence of these items could not be evaluated properly. As indicated above, the two versions of items were structurally equivalent when they had about the same correlation with the scales. As the table shows, most versions of the items had, indeed, correlations of the same order. For the evaluation of the difference between the correlations, Fisher's *z* transformation and the application for the evaluation of differences of *z* described by Hays (1974, pp. 661-665) were used. This application produced a *t* statistic and a corresponding *p* value that are presented in the last two columns of Table 4. For three items, namely, SF2, FC3, and FC6, the differences between the correlations were significant at the $p < .01$ level (two-tailed). For these items, the structural equivalence could not be accepted. For the other items the structural equivalence could be accepted.

THE CROSS-CULTURAL ITEM BIAS

The cross-cultural item bias was examined by comparing the mean scores of the individualistic versions (m_1) and the collectivist versions (m_2) of the items for the different groups. Note that there were more than two groups to compare (four ethnic groups and three levels of acculturation). The item bias was examined by the difference between the (standardized) Δe scores for the groups. The items were designed in such a way that Δe was more positive when pupils were more acculturated.

Table 5 shows the results of our analysis for one of the items (FC5) of the Family Cohesion scale. The individualistic version of the item was "When it is someone's birthday, we celebrate with the whole family," the collectivist version "When there is something to celebrate, the whole family participates." The two versions could be considered structurally equivalent because their correlations with the Family Cohesion scale were both significant and did not differ much (see Table 4). The individualistic version was designed to produce higher scores when pupils were more acculturated, the collectivist one was designed to do the same when pupils were less acculturated. As Table 5 shows, the individualistic version behaved as predicted: The scores were highest for pupils who were more acculturated. The same was true for Dutch pupils when compared with the other ethnic groups. As for the collectivist version, the differences between the versions were negligible when comparing levels of acculturation, although the scores of the Dutch pupils were lower than two of the three other ethnic groups.

Note that for four of the seven groups, the scores of the individualistic version were higher than the scores of the collectivist version. This could be explained by the lesser likelihood that the whole family participates in festivities when more types of festivities are included in the item. When a more specific list of festivities was included in the individualistic version of the item, the frequencies of the individualistic version could be lower than those of the collectivist one. This shows that it is not useful to interpret the values or signs of Δm or Δe scores separately. For the same reasons it is not useful to compare those scores between different items. The only meaningful way to use Δe scores is to compare these scores on the same item between groups in order to assess cross-cultural validity.

The crucial score differences were the mean differences of Δm (and Δe) between groups. The pattern of these differences was as predicted: The differences were higher when the pupils were more acculturated. Table 5 shows that the difference between Δm for the high-acculturated group and the low-acculturated group was 1.31. Because the pooled variance almost equaled 1 here, the standardized difference scores Δe were virtually the same. The difference between Δe for the high-acculturated group and the low-acculturated group was 1.26. This is a strong difference following the criteria of Cohen (see the end of the Theory section). There were also major differences between the Dutch and Turkish pupils. An analysis of variance shows that these differences were significant at the $p < .001$ level. Generally, it was concluded that Item FC5 was strongly biased. The reason for this could be that pupils from some ethnic minorities in general do not celebrate birthdays.

It is useful to evaluate the differences between the Δe scores when structural equivalence is likely. This was the case for Items SF1, SF3, SF4, SM1, SM2, SM3, SM4, FC1, FC5, FC7, and FC8. Table 6 shows the results for all these items. Note again that it is not useful to interpret the value or sign of Δe scores separately, because these scores depend directly on the distribution of these items. Therefore, the fact that the Δe scores of Items SM1 and SM2 are more negative than the Δe scores of the other items is not informative. It is only possible to interpret differences between Δe scores between groups. The Δe scores of almost all items were higher when the acculturation of the pupils was also higher. Only Item FC1 did not behave in line with expectations. The mean Δe score over all 11 items was .24 for the highest acculturation group, .02 for the mediate group, and –.39 for the lowest acculturation group.

The differences established between the ethnic groups were often as predicted. In 6 out of 11 cases the Δe scores were highest for adolescents with a Dutch background, lower for adolescents with a Surinamese background, and lowest for adolescents with Turkish or Moroccan background. The mean Δe score of all 11 items was .22 for adolescents with a Dutch background, .05 for adolescents with a Surinamese background, –.32 for adolescents with a Turkish background, and –.34 for adolescents with a Moroccan background. The pattern of the Δe scores was more obvious when comparing levels of acculturation than when comparing ethnic groups. The differences between the Δe scores of the groups were great for Items SM4 and FC5, whereas they ranged between minor and average for the other items. The differences between the scores of boys and girls were small. This indicates that most items were at least to some extent affected by cross-cultural item bias. An analysis of variance shows that major differences between Δe scores were also significant. Item FC5, for which structural equivalence was likely and the Δe differences were great, was the most explicit case of a cross-cultural bias.

TABLE 6

Cross-Cultural Bias of Items: The Δ*e* Scores of Individualistic and Collectivist Item Versions for Three Levels of Acculturation and Four Ethnic Groups

	Subpopulation by Degree of Acculturation				*Subpopulation by Ethnic Group*				
Item	*High*	*Medium*	*Low*	*Significance (p)*[a]	*Dutch*	*Surinamese*	*Turkish*	*Moroccan*	*Significance (p)*[a]
SF1	–0.17	–0.32	–0.46	.275	–0.24	–0.05	–0.39	–0.62	.374
SF3[b]	0.36	0.09	–0.21	.105	0.33	0.29	0.09	–0.11	.706
SF4[b]	0.05	0.06	–0.38	.313	–0.03	–0.72	–0.19	–0.40	.245
SM1	–0.28	–0.45	–0.56	.273	–0.30	–0.25	–0.54	–0.13	.679
SM2	–0.44	–0.74	–1.06	.003	–0.43	–0.90	–0.95	–1.10	.006
SM3[b]	0.56	0.46	–0.41	.006	0.61	0.81	–0.24	–0.05	.053
SM4[b]	0.18	–0.23	–1.53	.000	0.22	–0.57	–1.62	–1.24	.000
FC1	0.32	0.12	0.51	.266	0.24	0.37	0.30	0.06	.068
FC5	0.78	0.46	–0.48	.000	0.75	0.11	–0.51	–0.01	.000
FC7	0.25	0.10	–0.08	.217	0.30	0.03	0.16	–0.59	.007
FC8	1.03	0.69	0.36	.001	0.96	1.47	0.42	0.49	.013
Mean	0.24	0.02	–0.39		0.22	0.05	–0.32	–0.34	

a. Significance of the effect of the interaction between item version and group on the item (*F* test, analysis of variance).
b. The number of cases is between 570 and 604, which is about half the number of cases for the other items (see Table 4).

CONCLUSION AND DISCUSSION

In this article, a method has been presented for evaluating the presence and size of a cross-cultural bias in parental support items and family cohesion items in a questionnaire for adolescents. Each evaluated item had two versions, a more collectivist version and a more individualistic version, that is to say, a typical Dutch version. These versions were checked with regard to structural equivalence. The difference between the mean scores of these versions of an item, called Δe, gives an indication for the item bias of the item. In addition, the level of acculturation was also measured and used as an independent variable.

The structural equivalence was tested by comparing the correlations of the two item versions with the corresponding scale. Similar item-scale correlations were established for all but 6 items. This means that 10 items appeared to have structurally equivalent versions. One of these 10 items did not show an item bias. Although in many cases the bias was not significant, the other 9 items showed the expected patterns of a cross-cultural item bias. The biases differed in size. Nevertheless, for all items Δe was lower when acculturation was lower. For most items, Δe was highest for indigenous Dutch respondents, lowest for Turkish and Moroccan respondents, and intermediate for Surinamese respondents. Acculturation was highly correlated with ethnicity.

Although the questionnaire was tested among respondents of similar age and education, the questionnaire still showed a real item bias. This implies that standard questionnaires and tests, even when used within one country, must be treated with great care. It is obvious that the individualistic items are designed to be biased, and therefore it may be argued that it is not strange that biases were found. However, as Table 3 shows, the individualistic items appear to be quite normal: Initially, most items would not draw any attention in the questionnaire of a study. It is obvious that even slight differences in formulation may cause item biases. Pilot studies should be conducted to prevent an occurrence of such a bias.

All in all, researchers who want to compare ethnic groups or groups with various levels of acculturation should carry out a study on the cultural bias of their items. Such a study, preferably a pilot study, should contain two versions per item. These two versions must reflect two different cultural backgrounds and should be tested with regard to their structural equivalence. Subsequently, the standardized differences of group means of the versions (Δe) should be compared between the ethnic groups or between groups with different levels of acculturation. When Δe differs significantly between the groups, the item must be formulated again. When Δe is similar for the groups, both versions can be used.

NOTES

1. We presume here that designing a bias for a pilot study will succeed (the bias is in the direction intended) or will not succeed (there is no bias). When respondents misinterpret items, they are less sure about what is asked, and therefore they tend to score more neutral. It is possible that the mean real score on X is (semi)positive because most respondents are empathic. In this case, the mean scores on biased items will be lower. By design, the bias for item version I will be larger for population A (respondents with a visual handicap) than for population B, and the bias for item version II will be larger for B than for A. It is also possible that no bias emerges. In this case, we therefore conclude that:

$$m_e1(A) - m_e1(B) \geq 0 \textit{ and } m_e2(B) - m_e2(A) \geq 0. \quad (8a)$$

It is also possible that the mean real score of X is (semi)negative because most respondents are not empathic. In this case, the mean scores of biased items will be higher; that is, the biases are negative. By design, the bias for item

version I will be more negative for population A (respondents with a visual handicap) than for population B, and the bias for item version II will be more negative for B than for A. It is also possible that no bias emerges, and therefore:

$$m_\varepsilon 1(A) - m_\varepsilon 1(B) \leq 0 \textit{ and } m_\varepsilon 2(B) - m_\varepsilon 2(A) \leq 0. \quad (8b)$$

Now, consider again the situation that $\leq m(A) = \Delta m(B)$. From expression (6), it follows that

$$\Delta m(A) - \Delta m(B) = [m_\varepsilon 1(A) - m_\varepsilon 1(B)] - [m_\varepsilon 2(A) - m_\varepsilon 2(B)] = 0. \quad (9)$$

The effect of our design is either that (8a), both $m_\varepsilon 1(A) - m_\varepsilon 1(B) \geq 0$ and $m_\varepsilon 2(A) - m_\varepsilon 2(B) \leq 0$, or (8b), that both $m_\varepsilon 1(A) - m_\varepsilon 1(B) \leq 0$ and $m_\varepsilon 2(B) - m_\varepsilon 2(A) \geq 0$. In both cases, expression (9) can only be true when $m_\varepsilon 1(A) = m_\varepsilon 1(B)$ and $m_\varepsilon 2(A) = m_\varepsilon 2(B)$, which means, according to expression (2), that both versions I and II are not cross-culturally biased and could be used in the final study. It is concluded that a cross-cultural bias exists if and only if $\Delta m(A) - \Delta m(B) \geq 0$.

2. A two-way analysis of variance is appropriate here. The dependent variable is the score of the manipulated item. The independent variables are the versions of the item (I or II), the populations (A or B), and the interaction between them. The interaction term should be significant.

REFERENCES

Berry, J. W. (1990). Psychology of acculturation. In J. J. Berman (Ed.), *Cross-cultural perspectives, Nebraska Symposium on Motivation 1989* (pp. 201-234). Lincoln: University of Nebraska Press.

Berry, J. W., Trimble, J. E., & Olmedo, E. L. (1986). Assessment of acculturation. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 291-324). Beverly Hills, CA: Sage.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137-164). Beverly Hills, CA: Sage.

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

Hays, W. L. (1974). *Statistics for the social sciences* (2nd ed.). London: Holt, Rinehart and Winston.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage.

Hofstede, G. (1991). Empirical models of cultural differences. In N. Bleichrodt & P.J.D. Drenth (Eds.), *Contemporary issues in cross-cultural psychology—selected papers from a regional conference of the International Association for Cross-Cultural Psychology* (pp. 4-20). Amsterdam: Swets & Zeitlinger.

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

Janssens, J., Pels, T., Dekoviç, M, & Nijsten, C. (1999). Opvoedingsdoelen van autochtone and allochtone ouders [Educational goals of indigenous and foreign parents]. *Tijdschrift voor Orthopedagogiek*, *38*, 318-329.

Kortmann, F. (1990). Psychiatric case finding in Ethiopia: Shortcomings of the Self Reporting Questionnaire. *Culture, Medicine and Psychiatry*, *14*, 381-391.

Ott, L., Mendenhall, W., & Larson, R. F. (1978). *Statistics: A tool for the social sciences*. North Scituate, MA: Duxbury Press.

Pels, T. (Ed.) (1994). *Opvoeding in Chinese, Marokkaanse en Surinaams-Creoolse gezinnen* [Education in Chinese, Moroccan, and Surinamese-Creole families]. Rotterdam, The Netherlands: ISEO.

Poortinga, Y. H. (1995). Cultural bias in assessment: Historical and thematic issues. *European Journal of Psychological Assessment*, *11*, 140-146.

Poortinga, Y. H., & Malpass, R. M. (1986). Making inferences from cross-cultural data. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 17-46). Beverly Hills, CA: Sage.

Tanzer, N. K. (1991). A cross-cultural comparison of cognitive item structures and detecting cultural bias in nonverbal intelligence tests. In N. Bleichrodt & P.J.D. Drenth (Eds.), *Contemporary issues in cross-cultural psychology—selected papers from a regional conference of the International Association for Cross-Cultural Psychology* (pp. 428-436). Amsterdam: Swets & Zeitlinger.

Triandis, H. C. (1990). Cross-cultural studies of individualism and collectivism. In J. J. Berman (Ed.), *Cross-cultural perspectives, Nebraska Symposium on Motivation 1989* (pp. 41-134). Lincoln: University of Nebraska Press.

Triandis, H. C., McCusker, C., Betancourt, H., Iwao, S., Leung, K., Salazar, J. M., Setiadi, B., Sinha, J. B., Touzard, H., & Zaleski, Z. (1993). An etic-emic analysis of individualism and collectivism. *Journal of Cross-Cultural Psychology*, *24*, 366-383.

Van de Vijver, F.J.R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Newbury Park, CA: Sage.

Van de Vijver, F.J.R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*, 29-37.
Van Niekerk, M. (1993). Ethnic studies in The Netherlands: An outline of research issues. *Research Notes from The Netherlands*, *1*, 2-14.

Dianne van Hemert is a psychologist. She is currently a Ph.D. student in cross-cultural psychology at Tilburg University, The Netherlands. Her research interests include meta-analyses of cross-cultural differences and similarities in personality and emotion.

Chris Baerveldt is a sociologist. He worked from 1986 until 1992 as a researcher at the Research and Documentation Center of the Dutch Ministry of Justice. He is currently a senior researcher at Utrecht University, The Netherlands. His research interests include youth crime, social networks, social movements, and interethnic cultural strategies of adolescents.

Marjolijn Vermande is a developmental psychologist. She worked as a researcher at the Nijmegen Institute for Cognition and Information, and for the Netherlands Institute of Mental Health and Addiction. She is currently a lecturer/researcher at Utrecht University. Her research interests include children's social networks and relationships, assessment and taxonomy, and developmental psychopathology.