Cross-cultural generalization and universality

van de Vijver, F.J.R.; Poortinga, Y.H.

Published in:
Journal of Cross-Cultural Psychology

Publication date:
1982

Different meanings of the concept universality are distinguished and ordered according to the degree to which they are open to empirical control. Universality and specificity are considered as relative rather than absolute concepts. A relationship between the analysis of universality and the analysis of comparability or psychometric equivalence of data is established. An integrated approach to the analysis of universality and equivalence within the context of Generalizability Theory is outlined and illustrated with an example.

CROSS-CULTURAL GENERALIZATION AND UNIVERSALITY

FONS J.R. VAN DE VIJVER
YPE H. POORTINGA
Tilburg University
The Netherlands

Intergroup differences in behavior are a major raison d'être of cross-cultural psychology. In most studies in this field the differential effects of ecological or sociocultural factors are emphasized. The tradition of looking for differences appears to have been reinforced by cultural anthropologists, who long have searched for phenomena that are specific to a particular culture. This approach rests on the assumption that intergroup differences reflect real differences in psychological attributes. However, this interpretation has been challenged, notably on methodological grounds. It has been argued that differences in scores are likely to be caused by bias or lack of equivalence of measurement procedures, or other artefacts, rather than by the behavioral characteristics that were the object of study (e.g., Irvine & Carroll, 1980). Dissatisfaction with the theoretical status of cross-cultural differences is at least one of the reasons why the search for similarities in behavior across cultures and

---

AUTHORS' NOTE: The order of the names was determined at random.

Journal of Cross-Cultural Psychology, Vol. 13 No. 4, December 1982 387-408

for universal aspects of human behavior has been emphasized in recent publications (Jahoda, 1980; Lonner, 1980; Triandis, 1978; Warren, 1980).

A major problem with the concept of universality is that it does not have a precise meaning. Triandis defines a universal as "a psychological process or relationship which occurs in all cultures" (Triandis, 1978, p. 1). Another definition has been given by Jahoda (1981, p. 42), who considers "invariance across both cultures and methods" as the criterion for universality. These definitions differ in important respects. First, according to Traindis's definition, the occurence of a phenomenon in at least one individual in each culture is sufficient condition for the universality of a phenomenon. One can argue that this requirement is rather loose and that virtually all phenomena studied by psychologists are universal in this sense. In contrast, if a similar pattern of relationships in each culture is considered as a necessary condition, as Jahoda seems to maintain, the modal person in each population will have to display the behavior under consideration.

Second, the definitions differ with respect to methodological restrictions. Jahoda's definition requires that the same measurement procedure is used in all cultures. The requirement of using the same measuring device across cultures is not postulated by Triandis.

This indicates that various existing definitions of universality differ particularly in the extent to which they lend themselves to empirical scrutiny (see also Poortinga, 1982a). Four points can be identified along a dimension of "experimental rigor" or "strictness," which refer to four categories of universals: conceptual universals, functionally equivalent (weak) universals, metrically equivalent (strong) universals, and scalar equivalent (strict) universals.

CONCEPTUAL UNIVERSALS

This label refers to molar, theoretical concepts at a high level of abstraction. The most noteworthy example is the notion of

the "psychic unity of mankind" put forward by Boas (1911 [1965]) and later discussed by Kroeber (1948). Also in this class fall concepts like intelligence or adaptability (Biesheuvel, 1972), or sensotypes (Wober, 1966), as long as their meaning is not further specified in operational terms. The universality of concepts such as these cannot be refuted in (quasi-) experimental studies, since no empirical referents—that is, behaviours characteristic of the constructs—are supplied. For this reason the scientific status of conceptual universals in an explanatory framework can be strongly questioned. Although it is impossible to make psychologically meaningful comparisons between conceptual universals in different cultures, this need not mean that such broad labels should be abandoned altogether. We only submit that it is impossible to demonstrate empirically that, for example, Bushmen have adaped better or less well to the Kalahari desert than Eskimo to the Arctic environment, unless it is indicated what observable variables are considered relevant (availability of food supplies, exposure to a harsh climate, expected age, population growth, hunting skills) and how these should be measured.

FUNCTIONALLY EQUIVALENT OR WEAK UNIVERSALS

This category contains concepts for which empirical referents have been specified—although these may differ across cultures—and for which construct validity has been demonstrated in each culture (Cronbach & Meehl, 1955). Sometimes (partly) different measurement procedures are applied in different groups (e.g., Davidson, Jaccard, Triandis, Morales, & Diaz-Guerrero, 1976; Przeworski & Teune, 1970). From our perspective invariance of method is not an essential theoretical concern, as long as the validity of the measurements across cultures in respect of the same construct has been clearly established. We shall return to this point later on. When the same measurement techniques are used everywhere, the cross-cultural equivalence of a concept is often investigated by means

of correlational statistics, notably factor analysis (Brislin, Lonner, & Thorndike, 1973; Irvine & Carroll, 1980). An illustration of the factor-analytic approach is the three-dimensional structure underlying the affective meaning of words found in a large number of countries with the semantic differential technique (Osgood, May, & Miron, 1975).

It is clear that rigid testing of hypotheses in this context is possible, at least in principle, and that concepts with functional equivalence are universal in a qualitative, although not necessarily in a quantitative sense. This is easy to see if an example is considered. If the temperature is measured in two locations with different thermometers, one with a Celsius scale and the other with a Fahrenheit scale, meaningful comparisons of quantitative differences are impossible, although the same dimension is being measured.

METRICALLY EQUIVALENT OR STRONG UNIVERSALS

This class contains concepts that are measured in the same metric across cultures, although the scales may have a different origin in each culture. The Celsius scale and the Kelvin scale provide an analogy. While cross-cultural score comparisons of absolute magnitudes may be meaningless, intracultural score differences can be compared across cultures providing there is a common metric. For example, the difference between the score values of 1 and 11 in culture A is twice the difference between the score values of 20 and 25 in culture B. This idea has been applied in studies in which relative rather than absolute differences between cultures have been investigated (e.g., Cole, Gay, & Glick, 1968; Poortinga, 1971). A recent example is a study by Irvine and Reuning (1981) in which substantial intercultural differences in response latencies for encoding tasks were found. For four different tasks the results within each group come close to a theoretically predicted relationship stipulating quantitative differences between the tasks. Psychometric conditions for this kind of equivalence will be discussed in a later section.

SCALAR EQUIVALENT OR STRICT UNIVERSALS

Measurements of concepts in this category have to show an equal metric and equal scale origin in each culture. Practically speaking this will nearly always imply distributional identity across cultures.[1] Scalar equivalent data can be compared within and across cultures. Differences in means in the performance of culturally different groups on a presumably strictly psychometrically equivalent scale can appropriately be taken as falsifying the hypothesis that the construct is a strictly equivalent universal.

Only for very few concepts can strict equivalence be found at the present stage of cross-cultural psychology. Although this has not been established beyond all doubt, there is evidence that the speed of processing of simple auditory and visual stimuli as measured with simple reaction time experiments yields approximately the same average value across cultures (Jensen, 1980).

In sum, conceptual universals refer to molar, theoretical concepts without any reference to measurement scales; functionally equivalent universals are concepts for which empirical referents have been specified and that are measured in qualitatively the same way in each culture; metrically equivalent universals are concepts that have the same metric but not the same scale origin across cultures, and strictly equivalent universals have the same scale with the same origin in each culture. A close correspondence between the psychometric requirements mentioned here for conceptual, weak, strong, and strict universals, and the four levels of measurement in nominal, ordinal, interval, and ratio scales, as outlined by Stevens (1951), is obvious. In later writings (Torgerson, 1958) it has been emphasized that measurement in a proper sense requires the ordering of phenomena along a quantitative dimension, and this requires at least an ordinal scale. This is in line with our argument that conceptual universals as described here do not lend themselves to empirical investigation.

THE COMPARISON SCALE

Comparison across cultures implies that there is a common scale on which the comparison is made, apart from the observed score scales in the various cultures. This will be referred to as the comparison scale. Although it is possible to have a separate comparison scale, usually the observed score scale in one of the groups serves for this purpose. The relationship between any pair of scales can be represented in a transformation function. For such physical measurements as length or temperature the parameters of the transformation function between different measurement scales are known. In psychological measurement we usually do not have precise knowledge about the parameters of the transformation function between the metric of scales in different cultures. Therefore, the level of measurement of the comparison scale will often not be the same as that of the scales in separate groups. However, the measurement levels in the separate groups impose an upper limit on the measurement level of the comparison scale.

Relevant information about the comparison scales can be gained by studying relationships between observable variables within cultural groups. Irvine and Reuning's (1981) evidence can be taken to support the claim that across cultures the same processes are involved in encoding tasks. In principle such evidence can be obtained even when nonidentical measurement procedures are used in different cultures. This is the reason we stated earlier that in our view invariance of method is not an absolute condition for universality. On the other hand, it is very difficult to achieve an equal scale metric—not to speak of an equal scale origin—cross-culturally when formally different measurement procedures are applied. Therefore, a more direct way to gain information on the measurement level of the comparison scale is through the study of relationships between score variables obtained with formally identical measurement procedures across cultures.

Earlier work on this topic has led to a collection of loose or at best only vaguely connected psychometric conditions for

comparability or equivalence (see Poortinga, 1975, 1982b). Here a more coherent framework will be presented.

UNIVERSALITY VERSUS SPECIFICITY: A DICHOTOMY?

The current literature on universality, and to a lesser degree the literature on the emic-etic issue, seems to be based on the assumption that cross-cultural phenomena can be divided into universals and specifics, although opinions will differ about which phenomena belong to each category (e.g., Berry, 1969, 1972; Lonner, 1980). No framework has been developed for intermediate positions between universality and specificity. Nevertheless, it seems intuitively meaningful to consider the degree of invariance of data across cultural groups as a function of the similarity in cultural patterns or background variables between them.

It will be more difficult to obtain measurements with an equal origin and metric as the cultures involved become more dissimilar. Let us take a personality inventory, the MMPI, as an example. On the basis of a literature review Gynther (1972) has reported distinct differences between Blacks and Whites in the United States that he attributes partly to instrument-specific factors rather than to real differences. Apparently Gynther considers the MMPI scores for Whites living in different regions of the United States as—in terms of the previous section—strictly equivalent, while between Blacks and Whites only a lower level of equivalence is likely to hold. Although no pertinent data are available it seems likely that comparisons of MMPI scores between Americans and Bushmen would be virtually meaningless as they would lack equivalence at any empirically controllable level.

An approach that has sufficient flexibility for dealing with the kind of distinctions made above is the Generalizability theory as put forward by Cronbach and his associates (Cronbach, Rajaratnam, & Gleser, 1963; Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

Generalizability theory offers an important tool in cross-cultural comparisons as—and this is the essence of our argument—it makes a shift possible from the dichotomous concepts of universality versus specificity to the continuous concept of generalizability. By means of this theory it is possible to deal with different levels of equivalence, depending upon the specific groups from which results are compared. In this theory universality amounts to a high or maximum level of cross-cultural generalizability of measurements. A lower degree of cross-cultural generalizability indicates that we should be cautious with cross-cultural score comparison, and a complete absence of cross-cultural generalizability would be indicative for culturally specific aspects of behavior.

GENERALIZABILITY THEORY

In this section we shall briefly introduce Generalizability Theory, with special reference to the question of how it can be possible that a measurement refers to more universal or more specific aspects of behavior. A full treatment of the theory is given by Cronbach et al. (1963, 1972). A short introduction can be found in Van der Kamp (1976) or in Wiggins (1973). With the exception of unpublished work by Fyans (1977) little attention has been given to Generalizability Theory in cross-cultural psychology.

Generalizability Theory is essentially a liberalization of the reliability concept in classical test theory. In this theory reliability is closely associated with the idea of parallelism, since the reliability of a test is equal to its correlation with a parallel test. An example first mentioned by Guttman and discussed by Cronbach et al. (1972, p. 7) shows the ambiguity of the classical approach. A subject is given the task to write down as many words as possible that begin with the letter t. How should a test parallel to this task look? One may ask the subject to write down words that begin with the letter p or with the letter d, but it is equally plausible to ask for words that have the t as the second letter or that have the t as the final letter. If

correlations between the first task and each of the various parallel versions are computed and taken as an index of the reliability of the first test, different values may be found. It is obvious that none of the indexes can be considered as the reliable index.

In Generalizability Theory this ambiguity is resolved by introducing different domains or universes. The tasks in which the subject is asked to write down words that begin with t or p or d can be said to be sampled from the same universe, that is, the ability to generate words beginning with a fixed letter, whereas the ability to generate words with a fixed letter in any stipulated position is a broader universe. In general our data sets may be considered to be sampled from a variety of conditions or universes (particular stimulus contents, test formats, observers, occasions, cultures, subgroups, etc.). Each condition can be represented as a factor or facet in an analysis of variance model. For each systematic effect, either a main effect or an interaction, a generalizability coefficient may be estimated. Generalizability coefficients are analogous to reliability coefficients in classical test theory. They provide an estimate of the proportion of score variance that can be attributed to a certain source.

The most simple design in the investigation of cross-cultural equivalence occurs when the same measurement procedure is administered to groups of subjects in two or more cultures. In this design, labeled as design V-B by Cronbach et al. (1972, p. 38), stimuli (S) are crossed with persons (P), and persons are nested in cultures (C). The following effects can be distinguished:

S: Stimuli;
C: Cultures;
P, PC: Confounding of the main effect Persons and the Person by Culture interaction;
SC: Stimulus by Culture interaction;
PC, PSC, E: Confounding of the Person by Culture interaction, the Person by Stimulus by Culture interaction and an Error term (E).

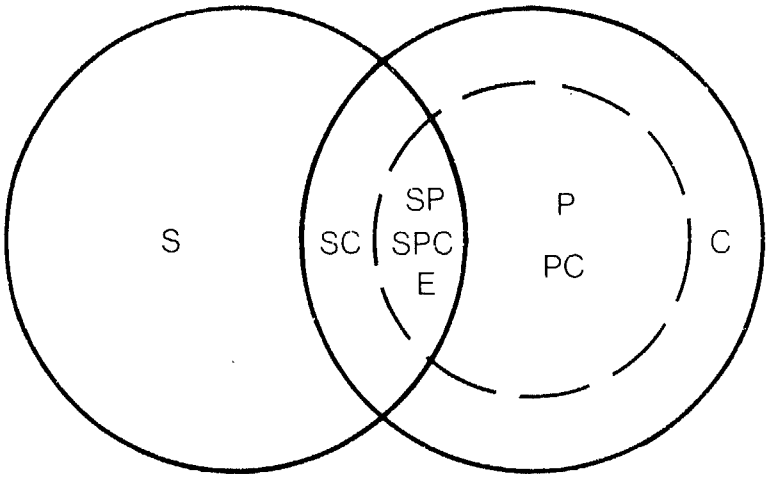


Figure 1: Schematic representation of variance components in a stimulus (S) by culture (C) design with persons (P) nested cultures.

In Figure 1 the composition of the scores is represented in a Venn diagram. In Generalizability Theory the variance components are estimated first in order to be able to estimate coefficients of generalizability.

The effect of each source, or combination of sources of systematic variance, can be estimated by means of a coefficient of generalizability. Particularly important for us (the reason for this emphasis choice will become clear in the next section) are (1) the coefficient estimating the stimulus by culture interaction, represented by $\hat{\rho}^2(\mu^{sc})$ and (2) the coefficient estimating the combined contribution of the main effect of culture and the stimulus by culture interaction, given by $\hat{\rho}^2(\mu^{c+sc})$.[2] (The hats above the symbols indicate that we are dealing with estimates.)

When generalizability coefficients have values close to zero this means that the corresponding sources do not contribute substantially to the score variance and thus do not form an impediment to quantitative comparisons across cultures.

THE DECISION PROCESS

How are decisions made in a generalizability framework about universality of the kinds distinguished earlier?

The procedure starts with the estimation of the variance components. This can be done with existing statistical packagers, for example, program P8V in the BMDP-series (Dixon, 1981), or by carrying out an analysis of variance (all random model) and then computing the components by hand with formulas as described in the Appendix.

The variance components of primary importance here are $\sigma^2(C)$, representing the main effect of culture and $\sigma^2(SC)$, indicating the interaction between stimulus nad culture. A substantial $\sigma^2(C)$ implies that consistent score differences across cultures are present. The psychological interpretation is equivocal since real cross-cultural differences cannot be separated from a uniform bias (Mellenbergh, 1982), which increases or decreases by the same amount the scores of all subjects in a culture.

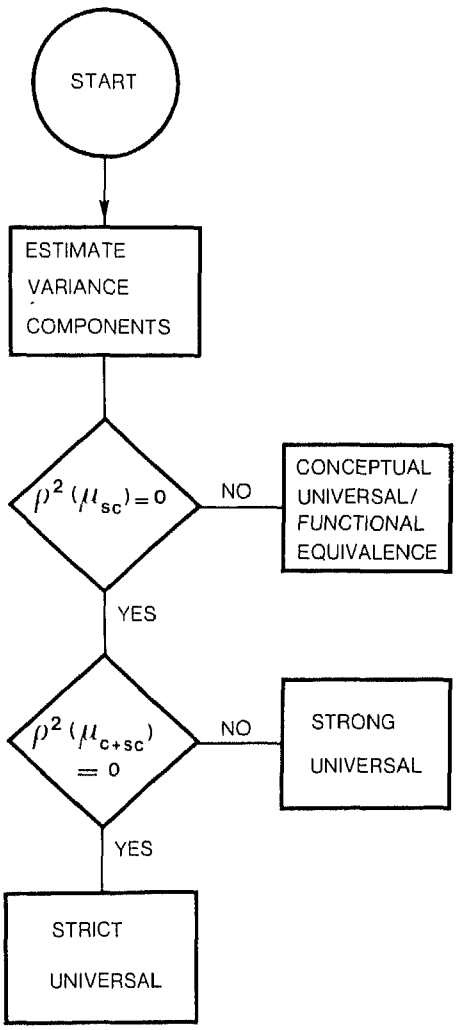
The second component of interest is the stimulus by culture interaction. When this component departs from zero this means that one or more stimuli are differentially influenced in one or more cultures. This effect may be "real", but the effect may also be due to factors such as bad translation of one or more items, statistical artifacts (Lord, 1977), heterogeneity of samples, lack of uniformity in administration procedure, and so on. It is obvious that the psychological interpretation of this interaction can be very difficult, and an adequate attribution of the cause of the interaction will require a closer inspection of the data, possibly resulting in a recommendation to extend the database or change the instruments.

The core of our approach consist of an inspection of the two generalizability coefficients $\hat{\rho}^2(_{\mu sc})$ and $\hat{\rho}^2(_{\mu^{c+sc}})$. These coefficients are defined in the appendix to this paper. Unfortunately the sampling distributions of these statistics are unknown, and hence no strict statistical criteria are available. Some information may be gained from the F-ratios. However,

Cronbach and others discourage this practice, because in their opinion, it is not the statistical significance level of a source as such that is important, but the size of the effect associated with that source. Some insight into the importance of sources of variance can be obtained by estimating the coefficients for various numbers of persons in the equations for the generalizability coefficients. When several thousand persons are tested, virtually each effect will be significant, although its practical importance may be limited (see Cleary & Hilton, 1968). There is another reason to be cautious with the interpretation of an F-ratio in this way. In most cross-cultural studies only a few cultures are sampled. If generalization to all existing cultures is intended, this implies that the component "culture" is usually poorly estimated.

In Figure 2 a flow chart of the decision procedure is given. After having estimated the variance components we first investigate whether $\hat{\rho}^2(\mu^{sc})$ differs substantially from zero. If this is the case, no meaningful quantitative comparison of scores is possible across cultures. A close inspection of the data may reveal the source of this interaction, and a reanalysis on a reduced data set can be useful. However, the elimination of items and of persons should be guided by criteria external to the analysis-of-variance table. Elimination motivated by an inspection of the residuals in each cell after removal of all main effects will result in chance capitalization, and as a consequence replication of the findings may become very unlikely. When $\hat{\rho}^2(\mu^{sc})$ differs from zero, two possibilities arise; the concept may be either a *conceptual universal,* or a *functionally equivalent* universal. The two will not be distinguished further here.

If, however, this coefficient may be considered to be negligible, our next step is to investigate $\hat{\rho}^2(\mu^{c+sc})$. Again, we try to see whether this coefficient differs from zero. If so, consistent cross-cultural differences exist and the particular construct is called a *strong universal,* characterized by the same metric but a different origin across cultures. (Whether the difference in origin is due to a uniform bias or is psychologically interpretable is of no concern here.) In this case



$\rho^2(\mu_{sc})$: GENERALIZABILITY COEFFICIENT; GENERALIZATION TO ALL COMPONENTS EXCEPT SC.

$\rho^2(\mu_{c+sc})$: ID. FOR C AND SC.

Figure 2: Flow chart of the decision process.

intracultural score differences can be compared meaningfully across cultures. However, if $\rho^2(\mu^{c+sc})$ is small, evidence is found for a *strict universal,* meaning that the scales have the same metric with the same origin across all cultures.

AN EXAMPLE

The data analyzed in this example were gathered by the second author. They are a small part of a study in which the universality of basic personality variables is investigated, specifically with respect to the "strength of the nervous system."

The samples consisted of Indian students, Dutch students, and Dutch army conscripts; all subjects were males and for this analysis each group was reduced to 32 subjects.

These 96 persons have answered a selection of 43 items forming the Strength of Excitation Scale in the third experimental edition of the Temperament Inventory (Strelau, 1972). Each question of this inventory has three response alternatives: affirmative (2 points), undecided (1 point), and negative (0 points). The scale had two parallel forms, the first with 22 items and the second with 21 items.

In the group of Indian students a split-half reliability of .72 was observed; for the Dutch students this was .78, and for the Dutch conscripts .75.

In Table 1 the results of the analysis of variance are given, together with the estimated variance components. The main effects for stimuli and persons—this latter confounded with the person-by-culture interaction—showed significant F ratios. There is no direct F ratio available in this design for testing the effect of the factor "culture"; therefore, a quasi-F ratio (Winer, 1971) was computed. The obtained F'(2, 129) was 2.50, which is strictly speaking not significant ($p = .09$). Since it is unclear to what extent the quasi-F ratio reflects the actual state of affairs, we should be cautious with statements about the (non) existence of culture effects. Finally, a highly significant item-by-culture interaction was observed. However, as noted earlier, these ratios do not have our primary interest.

TABLE 1
Results of the Analysis of Variance and the Estimated Components of Variance

Source	SS	df	MS	F	prob.	$\hat{\sigma}^2$
S	444.18	42	10.67	5.34	.00(+)	.0903
C	16.52	2	8.26	2.50[a]	.09	.0036
P, PC	172.11	93	1.85	3.37	.00(+)	.0303
SC	168.01	84	2.00	3.64	.00(+)	.0453
SP, SPC, E	2146.06	3906	0.55			.5494

a. Quasi F-ratio.

After having estimated the variance components our first step in the decision process (see Figure 2) involves the investigation of $\hat{\rho}^2(\mu^{sc})$, which is given in Table 2. The obtained unit sampling coefficient $\hat{\rho}^2(\mu^{sc}) = .08$. The degree of significance of this value is unknown. A close inspection of the data set and a comparison of generalizability coefficients with equivalent statistics with known distributions (e.g., Cronbach's α) led us to the conclusion that .05 may be a satisfactory, but possibly somewhat conservative, cutting point for the decision that a coefficient differs from zero. When we apply this rationale here this means that the Temperament Inventory does not yield quantitatively comparable scales for male samples in different cultures.

In subsequent analyses we tried to locate the sources of this inequivalence, mainly by restricting the universes of generlization. We started by taking the two Dutch groups together, thereby defining Dutch males from approximately 18 to 25 years as our universe. The value of $\hat{\rho}^2(\mu^{sc})$ was .02 (see Table 2), implying that the contribution of this component is considered negligible for all practical purposes. A value of .02 was also found for $\hat{\rho}^2(\mu^{c+sc})$. This test thus offers strictly universal, that is, quantitatively comparable results within the universe of young Dutch males.

As a next step we have further analysed the scores from the Indian and Dutch student samples together. For this new

TABLE 2
Estimated Generalizability Coefficients[a] (is = Indian students, ds = Dutch students, dc = Dutch army conscripts)

Groups	Unit/Full Sampling[b] $\hat{\rho}^2(\mu_{sc})$	Unit/Full Sampling[b] $\hat{\rho}^2(\mu_{c+sc})$	Comments
is, ds, dc	.08/.73	.08/.73	
ds, dc	.02/.42	.02/.40	only Dutch subjects
is, ds	.11/.80	.11/.80	only students
is, ds, dc	.08/.74	.08/.73	items with negative item-total correlations eliminated
is, ds, dc	.08/.73	.07/.71	1 of parallel forms (22 items)
dc	.07/.70	.14/.84	only Dutch conscripts, high vs. low scorers
dc	.01/.30	.01/.29	only Dutch conscripts; random split in two subgroups

a. Negative variance components were treated as zeros.
b. See Appendix.

universe, "male students of approximately 18 to 25 years," a value of .11 was found for $\hat{\rho}^2(\mu^{sc})$ (see Table 2), clearly indicating a lack of quantitative equivalence.

Until now, we have analysed the unscreened instrument. In an item analysis, not reported here, some negative item-total correlations were observed in one or more groups. After elimination of the eight items with negative correlations in at least one of the samples a value of .08 was found for $\hat{\rho}^2(\mu^{sc})$ (see Table 2). This implies that negative item-total correlations do not cause the lack of quantitative comparability.

During the item analysis it was observed that many items had rather high preference indexes and it was hypothesized that the lack of quantitative equivalence was caused by ceiling

ffects. In order to test this hypothesis one of the groups ıvestigated, the Dutch army conscripts, was split into two ıbgroups: a group with low scores, and one with high scores. 'or these subgroups $\hat{\rho}^2(\mu^{sc}) = .07$, in this case reflecting the item y subgroup interaction. For a random split of this group into vo subgroups a value of only .01 was observed. The value of)7 seems to be high enough to reject the hypothesis that item-y-subgroup interaction does not contribute to the score ariance. Subsequently, a similar pattern was observed in the ther groups.

After elimination of the items with means higher than 2.50 in t least one group, which was 58% of the items, a unit sampling oefficient of .05 was observed. This appears to confirm our npression that the observed lack of quantitative equivalence ; at least partly caused by ceiling effects. As a large number of :ems had to be removed before an acceptable value of the eneralizability coefficient was obtained, it is likely that dditional sources of inequivalence play a role.

In conclusion, we may state that the Temperament In-entory yields strictly universal scores for Dutch males from bout 18 to 25 years. However, we should be reluctant to ccept cross-cultural generalizations since substantial item-by-ulture interactions are observed, at least partly due to item eiling effects. In future studies with this scale it may be dvisable to replace the three-point item response scales by ive- or seven-point scales.

There is still another possibility. The test is composed of two ıarallel forms. When only one of these forms (with 22 items) vas analyzed, equal generalizability coefficients were observed .s were found for the total test (see Table 2). So, it seems worth onsidering to use only one of the test halves and to extend the est with a number of items with low preference indexes.

DISCUSSION

It is proposed here to substitute the concept of general-:ability for the culturally universal versus culturally specific

dichotomy in emprirical cross-cultural psychology, and to investigate this concept by means of the Generalizability Theory formulated by Cronbach et al. (1963, 1972). This shift offers a number of advantages. Universality is an absolute concept. It is meaningless to say that a phenomenon is more or less universal than another phenomenon. Generalizability is a relative concept; different degrees of generalizability can be meaningfully distinguished. Such a relative dimensional concept better meets the demands of cross-cultural psychology, since it is fairly obvious that most psychological phenomena of interest will be neither completely universal (i.e., equally frequently observable in all human beings), nor completely culturally specific (i.e., never occurring outside a particular culture).

Second, Generalizability Theory is a very flexible approach. In our example we analyzed a rather simple design, but extensions to much more complex designs can be readily made. Furthermore, all kinds of dependent variables may be used: test scores, self reports, observers' judgments, psychophysiological data, and so on.

Third, Generalizability Theory offers a coherent framework for the analysis of comparability of psychometric equivalence. Compared with earlier attempts (Poortinga, 1975, 1982b), the approach adopted here is straightforward. All computations are carried out within the framework of Generalizability Theory.

Finally, the use of generalizability coefficients forces us to define our universes. A generalizability coefficient always refers to a particular universe. When we define "culture" as our universe, as is often done, a "packaged" variable (see Whiting, 1976) is introduced, which usually offers a good prediction but an unsatisfactory explanation of score differences. It seems more informative to define universes with reference to a specific aspect or dimension of behavior, for example, quality of formal education, style of socialization, family structure, and so on (see also Segall, 1982).

The most important limitation of the approach as presented in the context of this article lies in its conservatism. The

occurrence of item-by-culture interactions and culture effects is ascribed to a lack of equivalence of measurement in the cultures involved. When consistent cultural differences are observed and the researcher is willing to attribute these to real cross-cultural differences, circumstantial evidence is needed to validate this choice and to rule out alternative hypotheses.

APPENDIX

Within the model described in the main text a score $X_{sp(c)}$ can be represented by

$$X_{sp(c)} = \mu + S_s + P_p, PC_{pc} + C_c + SC_{sc} + SP_{sp}, SPC_{spc}, E_{spc}$$

where:

μ is the overall mean;

S_s ($s = 1, \ldots, n_s$) is the main effect for stimuli;

P_p, PC_{pc} ($p = 1, \ldots, n_p$) is the confounded effect for the main effect persons and the person by culture interaction;

C_c ($c = 1, \ldots, n_c$) is the main effect culture;

SC_{sc} is the interaction between stimulus and culture;

$SP_{sp}, SPC_{spc}, E_{spc}$ is the confounding of the stimulus by person interaction, the stimulus by person by culture interaction and the error term (E).

The computation of the estimated generalizability coefficients starts with an analysis of variance. From the mean squares (MS) of this analysis variance components are estimated as follows:

$$\begin{aligned}
MS\,(SP, SPC, E) &= \hat{\sigma}^2(SP, SPC, E) \\
MS\,(SC) &= \hat{\sigma}^2(SP, SPC, E) + n_p\sigma^2(SC) \\
MS\,(P, PC) &= \hat{\sigma}^2(SP, SPC, E) + n_s\sigma^2(P, PC) \\
MS\,(S) &= \hat{\sigma}^2(SP, SPC, E) + n_p\sigma^2(SC) + n_p n_c\sigma^2(S) \\
MS\,(C) &= \hat{\sigma}^2(SP, SPC, E) + n_p\sigma^2(SC) + n_s\sigma^2(P, PC) \\
&\quad + n_s n_p\sigma^2(C)
\end{aligned}$$

in which $\sigma^2(C)$ represents the estimated variance component for the main effect culture, etc. The estimated generalizability coefficients $\rho^2(\mu_{sc})$ and $\rho^2(\mu_{c+sc})$ are computed as follows (cf. Cronbach et al., 1972):

$$\rho^2(\mu_{sc}) = \frac{\sigma^2(SC)}{\sigma^2(SC) + \sigma^2(SP, SPC, E)/n'_p}$$

$$\rho^2(\mu_{c+sc}) = \frac{\sigma^2(C) + \sigma^2(SC)}{\sigma^2(C) + \sigma^2(SC) + \sigma^2(P,PC)/n'_p + \sigma^2(SP, SPC, E)/n'_p}$$

where $n'_p = n_p$ for full sample estimates and $n'_p = 1$ for unit sampling.

Because a treatment of the differences between full sampling and unit sampling goes beyond this paper, it may be sufficient here to state that the use of unit sampling estimates is suggested in order to be able to compare generalizability coefficients from one study to another (cf. Golding, 1975).

NOTES

1. From a theoretical point of view this is not necessary. However, statistical tests for the condition that scales have an equal metric and an equal origin across cultures nearly always imply distributional identity. For example, many statistical techniques require normal distributions and homogeneity of variances.

2. It could be argued that the coefficient $\hat{\rho}^2(\mu_c)$ is of interest rather than $\hat{\rho}^2(\mu_{c+sc})$. The latter provides a more stringent condition for strict equivalence.

REFERENCES

Berry, J. W. On cross-cultural comparability. *International Journal of Psychology*, 1969, 4, 119-128.

Berry, J. W. Radical cultural relativism and the concept of intelligence. In L. J. Cronbach and P.J.D. Drenth (Eds.), *Mental Tests and Cultural Adaptation*. The Hague: Mouton, 1972.

Biesheuvel, S. Adaptability: Its measurement and determinants. In L. J. Cronbach and P.J.D. Drenth (Eds.), *Mental Tests and Cultural Adaptation*. The Hague: Mouton, 1972.

Boas, F. *The Mind of Primitive Man.* New York: Free Press, 1965 (originally published in 1911).

Brislin, R. W., Lonner, W. J., and Thorndike, R. W. *Cross-Cultural Research Methods.* New York: John Wiley, 1973.

Cleary, T. A., and Hilton, T. L. An investigation of item bias. *Educational and Psychological Measurement,* 1968, 28, 61-75.

Cole, M., Gay, G., and Glick, J. Some experimental studies of Kpelle quantitative behavior. *Psychonomic Monographs Supplement,* 1968, 10, (Whole No. 26), 173-190.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. *The Dependability of Behavioral Measurements.* New York: John Wiley, 1972.

Cronbach, L. J., and Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin,* 1955, 52, 281-302.

Cronbach, L. J., Rajaratnam, N., and Gleser, G. C. Theory of generalizability: A liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology,* 1963, 16, 137-163.

Davidson, A. R., Jaccard, J. R., Triandis, H. C., Morales, M. L., and Diaz-Guerrero, R. L. Cross-cultural model testing: Toward a solution of the etic-emic dilemma. *International Journal of Psychology,* 1976, 11, 1-14.

Dixon, W. J. *BMDP Statistical Software 1981.* Berkeley: University of California Press, 1981.

Fyans, L. J. New paradigm for cross-cultural psychological research. Paper presented at the meeting of the American Psychological Association, San Francisco, 1977.

Golding, S. L. Flies in the ointment: Methodological problems in the analysis of the percentage of variance due to persons and situations. *Psychological Bulletin,* 1975, 82, 278-288.

Gynther, M. D. White norms and Black M.M.P.I.: A prescription for discrimination? *Psychological Bulletin,* 1972, 78, 386-402.

Irvine, S. H., and Carroll, W. K. Testing assessment across cultures: Issues in methodology and theory. In H. C. Triandis and J. W. Berry (Eds.), *Handbook of Cross-Cultural Psychology,* Vol. 2. Boston: Allyn & Bacon, 1980.

Irvine, S. H., and Reuning, H. "Perceptual speed" and cognitive controls. *Journal of Cross-Cultural Psychology,* 1981, 12, 425-444.

Jahoda, G. Theoretical and systematic approaches in cross-cultural psychology. In H. C. Triandis and W. W. Lambert (Eds.), *Handbook of Cross-Cultural Pcychology,* Vol. 1. Boston: Allyn & Bacon, 1980.

Jahoda, G. Pictorial perception and the problem of universals. In B. Lloyd and J. Gay (Eds.), *Universals of Human Thought.* Cambridge: Cambridge University Press, 1981.

Jensen, A. R., *Bias in Mental Testing.* New York: Free Press, 1980.

Kamp van der, L.J.Th. Generalizability and educational measurement. In D.N.M. de Gruyter and L.J.Th. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement.* New York: John Wiley, 1976.

Kroeber, A. L. *Anthropology.* New York: Harcourt Brace, 1948.

Lonner, W. J. The search for psychological universals. In H. C. Triandis and W. W. Lambert (Eds.), *Handbook of Cross-Cultural Psychology,* Vol. 1. Boston: Allyn & Bacon, 1980.

Lord, F. M. A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic Problems in Cross-Cultural Psychology.* Lisse: Swets & Zeitlinger, 1977.

Mellenbergh, G. J. Conditional item bias methods. In S. H. Irvine and J. W. Berry (Eds.), *Human Assessment and Cultural Factors.* London: Academic, 1982.

Osgood, C. E., May, W. H., and Miron, M. S. *Cross-Cultural Universals of Affective Meaning.* Urbana: University of Illinois Press, 1975.

Poortinga, Y. H. Cross-cultural comparison of maximum performance tests. *Psychologia Africana Monograph Supplement,* 1971, 6.

Poortinga, Y. H. Limitations on intercultural comparison of psychological data. *Nederlands Tijdschrift voor de Psychologie,* 1975, 30, 23-39.

Poortinga, Y. H., The identification and measurement of psychological concepts across cultures. Paper presented at the WHO/ADAMHA International Conference on Classification and Diagnosis of Mental Disorders and Alcohol-and Drug-Related Problems. Copenhagen, 1982a.

Poortinga, Y. H. Psychometric approaches to intergroup comparison: the problem of equivalence. In S. H. Irvine and J. W. Berry (Eds.), *Human Assessment and Cultural Factors.* London: Academic, 1982b.

Przeworski, A., and Teune, H. *The Logic of Social Inquiry.* New York: John Wiley, 1970.

Segall, M. S. On the search for the independent variable in cross-cultural psychology. In S. H. Irvine and J. W. Berry (Eds.), *Human Assessment and Cultural Factors.* London: Academic, 1982.

Stevens, S. S. Mathematics, Measurement and Psychophysics. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology.* New York, John Wiley, 1951.

Strelau, J. A diagnosis of temperament by nonexperimental techniques. *Polish Psychological Bulletin,* 1972, 3, 97-103.

Torgerson, W. S. *Theory and Methods of Scaling.* New York: John Wiley, 1958.

Triandis, H. Some universals of social behavior. *Personality and Social Psychology Bulletin,* 1978, 4, 1-16.

Warren, N. Universals and plasticity, ontogeny and phylogeny: The resonance between culture and cognitive development. In J. Sants (Ed.), *Developmental Psychology and Society.* London: Mcmillan, 1980.

Whiting, B. The problem of the packaged variable. In K. F. Riegel and J. A. Meacham (Eds.), *The Developing Individual in a Changing World.* The Hague: Mouton, 1976.

Wiggins, J. S. *Personality and Prediction: Principles of Personality Assessment.* Reading, MA: Addison-Wesley, 1973.

Winer, B. J. *Statistical Principles in Experimental Design.* (2nd ed.). Tokyo: McGraw-Hill Kogakusha, 1971.

Wober, M. Sensotypes. *Journal of Social Psychology,* 1966, 70, 181-189.

Fons J. R. van de Vijver has a research position at Tilburg University, The Netherlands.

Ype H. Poortinga is on the staff of Tilburg University. He is an associate editor of the Journal of Cross-Cultural Psychology and is currently Secretary-General of the International Association for Cross-Cultural Psychology.