Available online at www.sciencedirect.com

ScienceDirect

Procedia Economics and Finance 20 (2015) 679 – 686

Procedia Economics and Finance

www.elsevier.com/locate/procedia

7th International Conference on Globalization and Higher Education in Economics and Business Administration, GEBA 2013

How reliable are measurement scales? External factors with indirect influence on reliability estimators

George Ursachi, Ioana Alexandra Horodnic, Adriana Zait*

University Alexandru Ioan Cuza Iasi, Bd.Carol I nr.11, 700504, Romania

Abstract

In social economic researches we often need to measure non-observable, latent variables. For this we use special research instruments, with uni and multi dimensional scales designed for measuring the constructs of interest. Validity and reliability of these scales are crucial and special tests have been developed in this respect. Reliability concerns often arise, due to external factors that can influence the power and significance of such tests. Even for standardized instruments variations are possible, and they could seriously affect research results. The purpose of the present study is to investigate if and how external factors could influence a largely used reliability estimator - Cronbach Alpha. Several scales commonly used in marketing researches were tested, using a bootstrapping technique. Results show that important differences in the values of Cronbach Alpha are possible due to indirect influence from external factors - respondents' age, gender, level of study, religiousness, rural/urban living, survey type and relevance of the research subject for the participants to the survey.

Peer-review under responsibility of the Faculty of Economics and Business Administration, Alexandru Ioan Cuza University of Iasi.

Keywords: Research methods, instruments, validity, scale reliability

1. Introduction

In social economic researches, the term "construct" is used for an ideal object that depends on a subject's mind (as opposed to a real object), an abstract idea or subject matter one wishes to understand, define and measure. These hypothetical constructs are not directly observable and are called latent variables. Different scales are used for

* Corresponding author. Tel.: 0040746013147; fax: 0040232201437.
E-mail address: azait@uaic.ro

Peer-review under responsibility of the Faculty of Economics and Business Administration, Alexandru Ioan Cuza University of Iasi.
doi:10.1016/S2212-5671(15)00123-9

measuring latent variables. Nunnaly and Bernstein (1994) stated there are two important issues of interest for science: developing measures for individual constructs and finding functional relationships between different constructs' elements. Research instruments need to simultaneously be valid and reliable. An instrument's reliability is given by its consistency in measuring a specific phenomenon; it supposes we get the same results for repeated measurements of the same phenomenon. Reliability is the extent to which an instrument will produce consistent results on similar subjects under similar conditions and can be assimilated with the precision of a certain measurement. It doesn't mean that the result is correct – or valid. A measurement instrument is valid when it really measures what it is supposed to measure (Peter, 1981; McGarland and Kimberly, 2005). Validity can be assimilated to the accuracy of a measurement or research instrument and is an indication of how sound a research is; it applies to both the design and the methods of a research. Validity implies reliability, but the reciprocal is not true; a valid measurement is reliable, but a reliable measurement isn't necessarily valid.

1.1. Validity and reliability types

Two broad categories of validity exist – external and internal. *External validity* checks if research results can be generalized or extrapolated for a whole population, for all similar situations or contexts, outside those in which the research took place. For the *internal validity*, related to the instrument itself, we speak about *content* (also called face, intrinsic or curricular) validity, present when the content of the research is related to the studied variables, has a logic; *criterion validity* (sometimes named concurrent validity) - how meaningful are the chosen research criteria comparing to other possible criteria; *construct validity* (or factorial validity) – which checks what underlying construct is actually being measured, and has three important parts - *convergent* validity (the degree to which two instruments designed to measure the same construct are related, convergence being found when the two analyzed instruments are highly correlated); *discriminant* validity – the degree to which two measures designed to measure similar, but conceptually different constructs are related, a low to moderate correlation being the proof of discriminant validity; *nomological* validity – the degree to which predictions from a formal theoretical network containing the analyzed concept are conformed, which means that constructs theoretically related are also empirically related, as well. (AERA, 1999; Diamantopoulos, 2005; McGarland and Kimberly, 2005; Peter, 1981)

Several types of reliability exist: internal consistency, split-half reliability, test-retest reliability and homogeneous reliability. (AERA 1999; Cronbach, 1951; Hulin, Netemeyer & Cudeck, 2001; Peterson, 1994) *Internal consistency reliability* is measured using the Cronbach Alpha coefficient, considered to indirectly indicate the degree to which a set of items measures a single unidimensional latent construct. This coefficient is a measure of the squared correlation between observed scores and true scores, reliability being thus measured in terms of the ratio of the true score variance to the observed score variance. A test is reliable if it minimizes the measurement error, so that the error is not highly correlated with the true score; at the same time, the relationship between the true score and the observed one has to be strong. The *split-half reliability* or equivalent reliability is also known as the Spearman Brown coefficient. In testing it, we randomly divide all items that pretend to measure the same construct into two sets, than we administer the entire instrument to a sample of people and calculate the total score for each randomly divided half. The split-half reliability estimate is simply the correlation between these two total scores. The *test-retest reliability* is also called stable reliability and checks what happens with the instrument in time - it assumes there are no substantial changes in the construct being measured between two different occasions. The *homogeneous reliability* is also labeled as inter-rater or inter-observer reliability and tells us if different investigators obtain the same results using the same instrument (are "calibrated". The percent of agreement between different raters or investigators is important, the correlation between the ratings giving us an estimate of the reliability or consistency between the raters.

1.2. Cronbach's Alpha problems

In marketing researches one of the most used reliability estimator is Cronbach's Alpha, introduced in 1951 by Cronbach, as a generalization of the KR-20 estimator created in 1937 by Kuder and Richardson. Cronbach Alpha is, due to its excessive use as well, a subject of controversies. Many authors question this estimator's power and are looking for alternative methods for testing a scale reliability (Bernardi, 1994; Christmann and Van Aelst, 2006; Green, Lissitz and Mulaik, 1977; Sitjma, 2009). Others are looking for ways of improving the quality and predictive

power of Cronbach Alpha, by identifying possible factors of influence. Most of the studies concentrated on internal factors - the number of items, the type of scale used or the length of the test (Kopalle and Lehmann, 1997; Mendoza et al., 2000; Norman, 1984; Osburn, 2000). More recent studies tried to investigate other possible influences, not related to the architecture of the instrument (Montag, 2008).

But why should we look for external influence factors? In computing the Cronbach's Alpha value we standardize, so variances of the estimator dissapear, the coefficient being calculated as a function of covariances. This means that there are no external factors with direct influence on the value of the estimator. However, in many cases different values of the estimator are obtained, for the same instrument, for the same type of scale and sample size. The logical conclusion is that these differences are due to external factors with indirect influence, which affect the relationship between internal constituents of the instrument. The aim of the present research is to identify such possible external factors.

An instrument's internal consistency is based on the correlation between different items of the same test. This correlation indicates if a number of items supposed to measure the same construct produce similar scores. For Cronbach's Alpha, computed with correlations between all pairs of items, internal consistency can vary between zero and one, although there are sometimes aberant negative values, as well (this implyies a negative average covariance among items, which could mean that while the true population covariances among items are positive, sampling error has produced a negative average covariance in a given sample of cases or that the items do not truly have positive covariances, and therefore may not form a single scale, they are not measuring the same thing). A general accepted rule is that α of 0.6-0.7 indicates an acceptable level of reliability, and 0.8 or greater a very good level. However, values higher than 0.95 are not necessarily good, since they might be an indication of redundance (Hulin, Netemeyer, and Cudeck, 2001). Although the ideal situation would be that all the items of a test measure the same latent variable, in many situations Cronbach Alpha can register high values even when the set of items actually measures different, independent variables (high values are not an indicator of scale unidimensionality) (Cortina, 1993; Cronbach, 1951; Green et al., 1977; Revelle, 1979; Schmitt, 1996; Zinbarg et al, 2006). In order to increase its' statistical power, Iacobucci and Duhachek (2003) suggest computing standard errors and judging Cronbach Alpha together with a confidence interval. While Cortina (1993) studied the Alpha coefficient through analitical comparisons, Peterson (1994) brings a huge contribution for marketing and consumer behavior research by investigating the quality of Alpha estimators from articles published in these fields. Collecting and analyizing hundreds of articles and thousands of coefficients, Peterson (1994) described a typical level of 0.77 (average for all articles). Kopalle and Lehman (1997) investigated the effect of eliminating the item with lower correlation, a technique that increases Alpha values and thus internal consistency. Indicators resulting from such a process could be super estimators with errors, when new Alpha values are computed for the same initial samples from which items were eliminated. Other authors have tried to increase the power of the Cronbach Alpha estimator through the elimination of its sensitivity to perturbating factors, such as outiliers (Christmann and Van Aelst, 2006).

2. Methodology

In this section we shortly describe our research methodology and reasons for selecting it. Besides the estimator's limits coming from its construction, as those indicated by Vehkalahti, Puntanen and Tarkkonen (2006), other issues appear – we can not obtain a distribution for Cronbach Alpha, so we can not apply statistical tests; we can not compare two instruments; we can not say anything about the significance of the differences obtained for two or more values of the estimator using separated samples; the value of the estimator is obtained post-factum, so we can not know the internal consistency of an instrument in advance, before using it. We had to find a way of creating a simulated distribution in order to facilitate testing and comparisons.

2.1. The bootstrapping technique

Bootstrapping was introduced by Efron in 1979. This procedure approximates the distribution of a statistical indicator using a Monte Carlo simulation, with multiple samples obtained from the empirical or the simulated distribution of the observed data (Cameron and Trivedi, 2005). The method consists of creating a distribution from a representative sample, by extracting new samples from the main one. Since the sample is representative and the

extraction is random and repeated, the distribution obtained through a large number of replications is very close to the one that would have been obtained if similar samples had been extracted from the whole population. Using this distribution we can compute standard errors, averages and quantiles and we can test differences for significance. Using the GAUSS program (with the generous help of Ruben Seiberlich, from Konstanz University), a distribution of Cronbach's Alpha was obtained, having as a starting point a representative sample for the investigated population.

2.2. Quantitative research

We used a questionnaire based survey, with three different instruments, applied to a total population of 900 persons, divided in 300 persons for each instrument. For each sample we ensured reprezentativity from the point of view of demographic variables of interest (external influence factors). 25% of the questionnaires were administered on-line and we also noted supplementary elements, such as level of noise and intimacy during questionnaire's completion.

The three instruments were: **A -** incorporates the scale need for cognition, with reported Alpha values between 0.62 and 0.89 and the scale need for uniqueness with Alpha values between 0.70 and 0.85 in previous studies; **B -** scale locus of control, with rather constant Alpha values of 0.70 – 0.71; **C –** incorporates the scales ease of use, with Alpha values between 0.83 and 0.88 and the scale anxiety - technological, with constant Alpha value of 0.93. Scales were translated to Romanian using the back translation procedure. Data were collected in the period January – May 2012.

2.3. Possible external factors

A number of 30 scales used in marketing research were investigated, paying attention to the reported Alpha's, type of survey, sample structure – especially demographic variables. Scales were chosen from Bruner (2009). We identified several groups of factors with potential influence on the estimator's values: cultural factors (level of education), living area, occupation, declared degree of religiousness); motivational factors (existence of a reward for questionnaire completion, relevance of the study for the investigated population, survey operator's gender); environmental factors (type of administration, level of noise, intimacy during questionnaire completion). Because external factors can not directly influence the value of the estimator, we can not build regression equations with these external factors; the constituent instrument's elements are different from one case to another, so we can not find general influence factors for the relationships between these elements. We can only find categories of factors with influence on a large number of cases and suggest possible factors depending on the characteristics of the research instrument.

In our quantitative research, the factors we analyzed had the following levels:
- Age: < 18 years, 18-25, 26-40, 41-65, and > 65 years old;
- *Gender:* Feminin and Masculin;
- *Living area:* Urban and Rural;
- *Education (studies):* under 8 classes, 8-10 classes, professional school, high school, postgraduate high school, bachelor, master degree, doctoral, post-doctoral studies;
- *Degree of religiousness* (declared): religious, somehow religious, not at all religious;
- *Noise level*: weak, strong and not-known (for the on-line participants);
- *Administration type*: face to face and online;
- *Existence of a reward*, with levels yes and no;
- *Relevance of the study for the respondent*, with levels relevant, non-relevant and not-known (for the on-line participants).

3. Main results

In tables 1 and 2 we present the main results just for the factors Age and Gender,
due to the length of the paper constraints. Similar analyses were made for all considered factors.

Table 1: Estimator's values on levels of age

Age levels (years)	**Alpha1 A** *need for knowledge* **(0.8804)**	**Alpha2 A** *need for uniquenes* **(0.7435)**	**Alpha B** *locus of control* **(0.6205)**	**Alpha1 C** *ease of use* **(0.9271)**	**Alpha2 C** *anxiety – technological* **(0.7222)**
< 18	0.8376	**0.1066**	0.6453	0.9967	0.8936
18-25	0.9061	0.7684	**0.4054**	0.8689	0.6752
26-40	0.8975	0.8750	**0.4610**	0.9055	**0.4170**
41-65	0.8899	0.7141	0.6858	0.8422	0.7677
> 65	0.8546	0.6932	0.6697	0.9735	0.7136

Table 2: Estimator's values on levels of the variable gender

Respondent gender	**Alpha1 A** *need for knowledge* **(0.8804)**	**Alpha2 A** *need for uniqueness* **(0.7435)**	**Alpha B** *locus of control* **(0.6205)**	**Alpha1 C** *ease of use* **(0.9271)**	**Alpha2 C** *anxiety – technological* **(0.7222)**
Feminin	0.8090	0.6824	0.5654	0.9352	0.7685
Masculin	0.9225	0.8011	0.6758	0.9208	0.6668

For the variable *Age (table 1)*, we found important variations for three of the scales (need for uniqueness, locus of control and anxiety – technological), with quite small values of Cronbach Alpha, indicating the instrument is not appropriate for those categories (values between 0.1 and 0.46).

For the variable *Gender (table 2),* the two groups – feminin and masculin – obtain different Alpha values for all the five cases, differences being really important in 4 of the 5 cases. We can not say anything about the sense of those differences - in 3 cases values are higher for the feminin group and in 2 cases for the masculin group.

The population from the urban area obtains, in 4 of the 5 cases, higher Alpha values than the rural population, the most important difference being obtained for the case (scale) anxiety – technological, for which the rural Alpha is very small. For the scale locus of control the difference is in the opposite sense, but not as high.

Differences registered between groups for the variable Level of Education (studies) are important for all the 5 cases (scales). For the locus of control scale we found the highest differences, with a very small value for the group post high school studies. Variations produced by the factor declared level of religiousness are not very strong but we noticed a tendency – Alpha values increased when the declared level of religiousness decreased. For the environmental noise variable, variations are small, but Alpha values are higher when the noise is weak. We did not find important differences for the type of administration of the questionnaire. The existence of a reward modified the Alpha values only for the scale anxiety-technological – values were over 0.7 for the rewarded group and under 0.7 for the nonrewarded one, showing that there is a potential problem.

We will continue by presenting results obtained using the program developed in Gauss. Again, results will be synthesized on factors (tables 3 and 4). We first obtained the distribution of the estimator for the whole sample, based on the simulation, than the Alpha value computed by the program, the standard error based on the simulation, the 5^{th} and 95^{th} quantile from the obtained distribution and the probability that the estimator's value is higher than 0.9 or 0.95 (for the whole sample). Then we compared these values with those for different groups. We tested only scales with a number of three items or less, due to the programming constraints. A much more complex program is needed for a higher number of items; however, most problems usually arise for scales with a small number of items, so this is not necessarily a strong limitation of the study.

Table 3: Values of the estimator for Age (scale C1 ease of use)

Point Estimate or Original Crombachs Alpha: 0.92713089
Bootstrapped Standard Error: 0.011683411
5% Quantile: 0.90592119 *95% Quantile: 0.94366057*
Monte Carlo average 0.92640031
Probability of being above 0.90000000 : 0.984000000 = 98,4% *Probability of being above 0.95000000 : 0.013000000* = 1,3%

When we compared values for different age segments, we noticed that for the group 18-25 years the Alpha values is ***0.86894442***, similar with that from SPSS, of 0.8689. The standard error for the simulated distribution for this group is of *0.031348813,* the 5% quantile is *0.80305927 and* the 95% quantile is *0.90949584*, which means that 90% of distribution's values are between these values. The average of the simulated distribution is of *0.86773665*, close to the value computed for the estimator. For the obtained distribution there is a probability of *71,2 %* that the estimator value is above 0,85. From a statistical point of view we can state, with a 99% probability, that the value of Alpha for the segment 18-25 years is different from 0,9, because the value of the statistical t test is *9.9064625,* higher than the table value 2.576 (p=0.005). A similar judgment was made for the other age segments showing important differences for the Cronbach Alpha estimator for different groups inside the sample comparing to the whole sample.

Table 4: Values of the estimator for Level of Education, High School studies

Point Estimate or Original Crombachs Alpha: ***0.88428603***
Bootstrapped Standard Error: 0.019007209
5% Quantile: 0.84861202 *95% Quantile: 0.91404290*
Monte Carlo average: 0.88154058
Probability of being above 0.90000000 : 0.170000000 = 17,0%

When we compared the situation of this group (table 4) with that of the whole sample and than with that of the group post high school studies, we noticed significant differences. The calculated value of Alpha for the segment high school is ***0.88428603***, similar with the SPSS value of 0.8843. The standard error of the simulated distribution is *0.019007209*, value used for applying tests to the estimator. The 5% quantile is *0.84861202 and* the 95% quantile is *0.91404290*, which means 90% of the distribution's values are between these values. There is a 17% probability that the estimator is above 0.9. For the segment post high school studies, the calculated values of Alpha for the segment high school is ***0.73554604***, similar with the SPSS value of 0,7355. The standard error of the simulated distribution for this group of studies is *0.082700012*, value used for applying tests to the estimator. The 5% quantile is *0.60361777 and* the 95% quantile 95% is *0.88542422*, which means 90% of the distribution's values are between these values. There is a 63% probability that the estimator is above 0.7. From a statistical point of view, with a 99% probability, the value of Alpha for the post high school group is different from 0.7, since the statistical t test value of 4.2981903 is higher than the table value of 2.575 (p=0.005). Also, from a statistical point of view, with a 99% probability, the value of Alpha for the high school group is different from the value of Alpha for the post high school group, since the statistical t test value of 17.528487 is higher than the table value of 2.576 (p=0.005). Following the same judgment we noticed that differences between the other groups of studies are also significant.

4. Conclusions

After running our bootstrap analyses for all the factors, we discovered 9 variables as external factors to the tested research instrument (scale C1 – ease of use) that have significant influences on the values of the Cronbach Alpha estimator. The intensity of the influences is different from one scale to another and also the sense of these differences can be different. For one factor – level of noise – the sense is the same for all the scales – a higher noise

determines lower Alpha values. Other 3 factors obtained the same sense in 4 of the 5 scales, and these factors are declared degree of religiousness (higher values for less religious people), living area (higher values for urban area) and type of administration (higher values for face to face).

The Alpha Cronbach value for the scale ease of use is different for different age groups, statistical tests showing that the scale has not the same reliability for the segments 18-25 years, 26-40 years and 41-65 years old. Also, the estimator's values for the scale ease of use are different for the feminin and masculin groups, for different levels of studies and for different degrees of interest – relevance of the study to the respondents.

Our research - from which just partial results are presented, signals potential reliability issues for some of the most used scales in marketing research. When populations are very heterogeneous, these reliability issues could be worse, especially for instruments with lower reliability estimator's values. In these particular cases the research instruments are not appropriate for some categories inside the whole population. Although further tests are necessary, especially for scales with a larger number of items, some important implications can be noticed at this point:
- the large differences identified for the Alpha estimator's values for sub-samples of the analyzed sample suggest we need to be cautious when we rely on instruments whose values are at the bottom limit of acceptability;
- reliable instruments need to be tested and adapted to the specific analyzed population, even if they were previously validated on other populations;
- supplementary reliability checking is necessary for instruments applied for very heterogeneous populations.

Certain limits of this study have to be considered. In order to increase the precision, larger samples – at least 10 times larger – would be needed, in order to have better reprezentativity within sub-samples. Also, testing more instruments, grouped on categories, would offer a clearer, bigger image. Developing a GAUSS program that could be applied to scales with more than 3 items would also be useful, although it is complex and time consuming. Other factors could also be tested, considering what specialists lately discovered in sociological researches (Henrich, Heine and Norenzayan, 2010), looking for potential differences caused by the fact that most of the scales were tested on "weird" people – persons from Western, Educated, Industrialized, Rich, and Democratic countries.

The main contribution of this study, together with the bootstrapping procedure developed, is the fact we signal potential problems for existing coefficients evaluating scale reliability and draw attention to a necessary cautious treatment of external factors with indirect influence on the Cronbach Alpha reliability estimators.

References

AERA. (1999). Standards for educational and psychological testing. *American Educational Research Association.*

Bernardi, R. A. (1994). Validating research when Cronbach's alpha is below 0.70: A methodological procedure. *Educational and Psychological Measurement, Vol. 54,* 766-775.

Bruner, G. (2009). *Marketing Scales Handbook, vol. 5.* Carbondale, Illinois: GCBII Productions.

Cameron A. C. and Trivedi P. K. (2005). *Microeconometrics. Methods and Applications.* New York: Cambridge University Press.

Christmann, A. and Van Aelst, S. (2006). Robust Estimation of Cronbach's Alpha. *Journal of Multivariate Analysis, Vol. 97,* 1661-1674.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology,* 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, Vol.16,* 297-334.

Cronbach, L. J. (1988). Internal consistency of tests: Analyses old and new. *Psychometrika,* 53 - 63.

Diamantopoulos, A. (2005). Procedure for scale development in marketing. A comment. *International Journal of Research in Marketing,* 1-9.

Green, B. F. (1981). A primer of testing. *American Psychologist nr.36,* 1001–1011.

Green, S. B., Lissitz, R.W., and Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement nr.37,* 827–838.

Henrich, Joseph, Heine, Steven J. and Norenzayan, Ara (2010). The weirdest people in the world? *Behavioral and brain sciences, vol.33,* 61-135.

Hulin, C., Netemeyer, R., and Cudeck, R. (2001). Can a Reliability Coefficient Be Too High? *Journal of Consumer Psychology, Vol. 10, Nr. 1,* 55-58.

Iacobucci, D. and Duhachek, A. (2003). Advancing Alpha: Measuring Reliability with Confidence. *Journal of Consumer Psychology, Vol. 13,* 478-487.

Kopalle, P. K., and Lehmann, D. R. (1997). Alpha inflation? The impact of eliminating scale items on Cronbach's alpha. *Organizational Behavior and Human Decision Processes, Vol.70,* 189-197.

McGarland R., and Kimberly, K.-L. (2005). Content Validity. *Encyclopedia of Social Measurement,* 495-498.

Mendoza, J.L., Stafford, K.L., and Stauffer, J.M. (2000). Large sample confidence intervals for validity and reliability coefficients. *Psychological Methods,* 356-369.

Montag, C. (2008). Does Speed in Completing an Online Questionnaire Have an Influence on Its Reliability? *Cyberpsychology & Behavior, Vol. 11,* 719-721.

Norman, C. (1984). An Improved Internal Consistency Reliability Estimate. *Journal of Educational Statistics, Vol. 9,* 151-161.

Nunnally, J. C. and Bernstein, I. H. (1994). Psychometric Theory ed.3. *New York: McGraw Hill.*

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods,* 343-355.

Peter, P. J. (1981). Construct Validity: A Review of Basic Issues and Marketing Practices. *Journal of Marketing Research, Vol. 18, Nr. 2,* 133-145.

Peterson, R. A. (1994). A Meta-analysis of Cronbach's Coefficient Alpha. *Journal of Consumer Research, Vol.21,* 381-391.

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research nr.14,* 57–74.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment nr.8,* 350–353.

Sitjma, K. (2009). On the use, misuse and the very limited usefulness of Cronbach's Alpha. *Psychometrika, Vol.74,* 107–120.

Vehkalahti, K., Puntanen, S. and Tarkkonen, L. (2006). Estimation of Reliability: a better alternative for Cronbach's Alpha. *Elsevier Science,* 1-19.

Zinbarg, R., Yovel, I., Revelle, W. and McDonald, R. (2006). Estimating generalizability to a universe of indicators that all have an attribute in common: A comparison of estimators. *Applied Psychological Measurement nr. 30,* 121–144.