International Journal of Psychological Research, 2010. Vol. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

Fischer, R., Milfont, T. L., (2010). Standardization in psychological research. *International Journal of Psychological Research, 3* (1), 88-96.

Standardization in psychological research.

Acerca de la estandarización en la investigación psicológica.

Ronald Fischer
Victoria University of Wellington
Taciano L. Milfont
Victoria University of Wellington

ABSTRACT

The term standardization has been used in a number of different ways in psychological research, mainly in relation to standardization of procedure, standardization of interpretation and standardization of scores. The current paper will discuss the standardization of scores in more detail. Standardization of scores is a common praxis in settings where researchers are concerned with different response styles, issues of faking or social desirability. In these contexts, scores are transformed to increase validity prior to data analysis. In this paper, we will outline a broad taxonomy of standardization methods, will discuss when and how scores can be standardized, and what statistical tests are available after the transformation. Simple step-by-step procedures and examples of syntax files for SPSS are provided. Applications for personality, organizational and cross-cultural psychology will be discussed. Limitations of these techniques are discussed, especially in terms of theoretical interpretation of the transformed scores and use of such scores with multivariate statistics.

Key words: standardization, score transformations, normal distribution, z scores, response styles, social desirability, culture

RESUMEN

El término de estandarización ha sido usado en varias formas en la investigación psicológica, principalmente en relación con la estandarización de procedimientos, estandarización de la interpretación, y estandarización de puntuaciones. El presente artículo se enfoca en el último tipo de estandarización. La estandarización de puntuaciones es una práctica común en escenarios donde los investigadores se interesan por ejemplo en diferentes estilos de respuestas. En este contexto, las puntuaciones son transformadas para incrementar su validez previa al análisis de datos. En este artículo delineamos una amplia gama de métodos de estandarización, discutimos cuándo y cómo las puntuaciones pueden ser estandarizadas, y qué pruebas estadísticas existen para tratar datos transformados. Se proveen procedimientos paso a paso y ejemplos en SPSS. Se discuten las aplicaciones de este procedimiento en investigación en personalidad, psicología organizacional y psicología transcultural. También se discuten las limitaciones de estos métodos, especialmente en relación con la interpretación de puntuaciones transformadas y su uso en estadística multivariada.

Palabras clave: Estandarización, Transformación de valores, Distribución Normal, valores z, estilos de respuesta, aceptación social, cultura

Article received/Artículo recibido: December 15, 2009/Diciembre 15, 2009, Article accepted/Artículo aceptado: March 15, 2009/Marzo 15/2009
Dirección correspondencia/Mail Address:
Ronald Fischer, School of Psychology, Victoria University of Wellington. Email: Ronald.Fischer@vuw.ac.nz
Taciano L. Milfont, School of Psychology, Victoria University of Wellington Email: Taciano.Milfont@vuw.ac.nz

INTERNATIONAL JOURNAL OF PSYCHOLOGICAL RESEARCH esta incluida en PSERINFO, CENTRO DE INFORMACION PSICOLOGICA DE COLOMBIA, OPEN JOURNAL SYSTEM, BIBLIOTECA VIRTUAL DE PSICOLOGIA (ULAPSY-BIREME), DIALNET y GOOGLE SCHOLARS. Algunos de sus articulos aparecen en SOCIAL SCIENCE RESEARCH NETWORK y está en proceso de inclusion en diversas fuentes y bases de datos internacionales.
INTERNATIONAL JOURNAL OF PSYCHOLOGICAL RESEARCH is included in PSERINFO, CENTRO DE INFORMACIÓN PSICOLÓGICA DE COLOMBIA, OPEN JOURNAL SYSTEM, BIBLIOTECA VIRTUAL DE PSICOLOGIA (ULAPSY-BIREME), DIALNET and GOOGLE SCHOLARS. Some of its articles are in SOCIAL SCIENCE RESEARCH NETWORK, and it is in the process of inclusion in a variety of sources and international databases.

International Journal of Psychological Research, 2010. Vol. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

Fischer, R., Milfont, T. L., (2010). Standardization in psychological research. *International Journal of Psychological Research, 3* (1), 88-96.

Standardization is a term widely used in psychological research, but that has a number of different meanings. Here we discuss three common usages of the term, namely standardization of procedure, standardization of interpretation and standardization of scores (Murphy & Davidshofer, 1994). We then discuss issues related to score standardization in more detail, since these procedures are fairly common in psychology, especially in applied areas of psychological research, and since they have sometimes unknown, but wide-ranging implications for descriptive and inferential statistics.

STANDARDIZATION OF PROCEDURE

Standardization of procedure is basic ‘experimental control’. For this reason this type of standardization is of supreme importance in any kind of research design (e.g., Harris, in press). ‘Experimental control’ is a basic requirement for any psychological test or experimental procedure. Typically this includes the standardization of instructions, administration (including manipulation) and measurement of variables of theoretical interest. Clear instructions appropriate for the particular population need to be provided. If verbal instructions are necessary, consideration should be given to the rate of speaking, tone of voice, inflections, facial and bodily expressions or pauses. A good example of how imprecision in administration can lead to biased results is evident in the testing literature. For example, if the experimenter smiles when a correct answer was given or slightly pauses prior to reading out the correct answer will influence the testing behaviour of the participants (and may increase or decrease test scores artificially). The order of presentation of test material or experimental manipulations needs to be identical, completely random or counterbalanced between participants, trials and conditions. Many experimental and testing designs require that time constraints are specified. It is also important to anticipate questions by participants and to develop guidelines for how to handle questions.

In essence, this type of standardization tries to reduce the influence of any extraneous variable on the test or experimental performance of participants. If procedures are not standardized, this will affect reliability and internal validity and result in biased findings. Standardization of procedure is often used in the context of test development (e.g., test instructions, item order, time limits) than in experimental design. In experimental designs, these issues are often discussed under the headings of internal and external validity.

STANDARDIZATION OF INTERPRETATION

This type of standardization refers to the standardized interpretation of obtained scores, often in the context of psychological test administration and interpretation (Murphy & Davidshofer, 1994). Scores of psychological tests are often not interpreted in their raw form, but against so-called norms. Psychological tests typically have no predetermined standards against which performance of individuals or groups of individuals can be evaluated. As a consequence, test scores are compared to some norm that was obtained by applying the same test in a sample supposed to represent the population. Norms can be seen as the typical, normal or average performance within a population. For example, if the typical 6-year child completes 5 tasks out of 20 correctly, this would constitute the norm for 6-year olds. Without this knowledge of the ‘average’ child, the raw score of 5 out of 20, or 25% correct answers, would be meaningless. Test scores in psychology can take any number of forms, including correct answers (absolute number of percentage), errors committed, central tendency on an attitude or personality inventory, reaction times, or any other objectively measurable performance that is indicative of a psychological construct. Norms can also be derived in a number of ways. The example above was an age-norm. Other norms include population norms (gender, group) or grade norms. These norms are calculated by administering the same test to a large sample that is representative of the population of interest. The representativeness of sample is a major factor in the judgment of norms. A mental ability test designed for gifted adolescents (to make finer distinctions between highly talented individuals than would be possible with tests designed for the whole range of abilities) should be normed in a large sample of gifted adolescents. To derive norms based on a sample of average students, school dropouts or students undergoing psychiatric treatment would be meaningless.

Despite the appeal of norms for ease of interpretation, the utility and applicability of norms is hotly debated, especially in non-academic circles. Norms are common in many applied contexts (e.g., in mental health, educational and work settings), where decisions need to be made about individuals falling above or below some particular normative criterion (e.g., for selecting individuals for therapy, students for university entry or hiring new employees). One of the big issues related to the utility and applicability of norms is that norms are population dependent. In many societies today, various minority groups exist. These groups often are disadvantaged along a number of important variables, including access to education and professional development. Applying norms based on majority group individuals to minority group individuals is likely to yield inappropriate conclusions, especially if the performance in the test is contingent upon disadvantages faced by the minority group (e.g., education dependency of mental tests). A second concern is that norms are often context specific or not stable over time. The Flynn effect is probably the most well-documented example for the temporal instability of norms. It refers to the phenomenon that IQ scores have been steadily

increasing over the decades (Brouwers, van de Vijver & van Hemert, 2009), which requires constant re-norming of mental ability tests. Using outdated norms will lead to invalid inferences about individuals.

In summary, the interpretation of test scores can be highly misleading if inappropriate or outdated norms are used. Some researchers have therefore called for avoiding global norms and argue for local norms (norms specific for a local population) (see for example, http://ipip.ori.org/newNorms.htm, last accessed November 6, 2009). Some went even further and called for abandoning testing altogether (it is noteworthy that these voices are stronger outside the academic psychological community than within, e.g., see http://www.fairtest.org/whats-wrong-standardized-tests) or advocate a person-centred view (e.g., Hofstee, 2008; Lamiell, 2007). Our reaction would be that although these critiques are valid, abandoning norms for comparative purposes would be immature in practical settings. Most importantly, it would open the doors for subjective and biased decision-making if no standards for interpretation are applicable. Norms were designed to avoid exactly these subjectivities in interpretation and to provide a more transparent, valid, fair and reliable way of conducting psychological testing.

STANDARDIZATION OF SCORES

The final use of the term discussed here is in relation to the form of individuals or group scores. Scores are often not interpreted in their raw form, but are converted to some other metric. There is a link between stardardization of interpretation discussed above and stardardization of scores, particularly since norm scores are sometimes expressed in z-scores (sometimes called standardized scores). Here, raw scores are expressed in units that indicate the position of an individual relative to the distribution of scores in his or her group. A score of 0 would indicate that the individual has a score that is exactly at the mean of the group, whereas a score of 1.0 would mean that the individual score is one standard deviation above the mean for this group and a score of -1 indicates a position of one standard deviation below the mean. Similar to scores based on population norms, z-scores allow normative interpretations. The difference is that the interpretation is sample specific, the position of each individual is evaluated against all others in the sample. The ad-hoc sampling of participants in psychological studies (typically involving relatively small samples of university undergraduate students) together with often non-normal distributions of psychological variables in small samples does not allow a meaningful comparison of z-scores across studies and samples. Cohen, Cohen, Aiken and West (1999) proposed one alternative to overcome some of these problems. They suggested the use of Percent of Maximum Possible (POMP) Scores which express raw scores in terms of the maximum possible score. Any score can be converted into POMP scores by taking the raw score minus the minimum score and then divide it by the possible scoring range. If test scores had binary response options, this would be equivalent to the percentage of correct answers. If multiplied by 100, the converted scores range between 0 and 100. This scoring method effectively standardizes the scores, allowing comparisons across alternative scoring methods, populations and instruments. Box 1 shows the calculation of POPM scores for SPSS, using SPSS syntax.

Box 1: Calculation of POMP scores

Example of SPSS syntax (to be copied and pasted into a SPSS syntax window):

Compute *POMP*=((*variable_name* - *minimum_score*)/(*maximum_ score* ï *minimum_ score*))* 100.

For variable name include the variable name in your SPSS spreadsheet. For minimum score include the lowest possible score, on a Likert scale from 1-5 this is 1, on a semantic differential type scale ranging from -3 to +3, this is a -3. Similarly, for maximum score use the highest possible score, in our example this would be either 5 or +3. The multiplication with 100 is not necessary, it will create scores that vary between 0 and 100 (hence the name ôPercentage of maximum possible scoresõ). Items in italics need to be changed.

Furthermore, POMP scores have the advantage of conveying immediate meaning. Psychological tests have typically arbitrary scales. Tests measuring the same construct have often different answer scales and it is not clear how a score of 3 on a scale from 1 to 4 compares with a score of 3 on a scale from 0 to 4. This arbitrariness of scaling hinders efforts to build a more cumulative science. The use of POMP scores does not overcome the issues of interpretation, since they do not address the relative meaning of scores compared to some real or statistical norm. A score of 30 only indicates that the individual achieved a score of 30, but could have obtained a score of 100 in a test. This does not tell us whether 30 is relatively high or low in comparison to his or her peers.

STANDARDIZATION TO ADJUST FOR RESPONSE STYLES

In the following we will discuss one important application of stardardization of scores in some applied areas of psychology, including personality measurement as well as in cross-cultural psychology. The issue is standardization of scores based on subjective evaluations of attitudes, personality, values, beliefs or some other psychological construct measured using some form of rating scale. As before, it deals with the problem of score

comparability, but here the concern is the presence of some response style or bias that systematically distorts the observed raw scores. Note the difference. The previous concerns of standardization were with the interpretation of the scores, but the raw scores were taken as valid and the only problem was the interpretation of each score in relation to the either norm population or the scale used for measuring the construct (i.e., issues related to stardardization of interpretation). Here the problem is that the raw score is seen as biased and needs to be corrected (standardized) to reveal its ótrueô score.

There are two particular types of response styles that have been widely discussed in the literature (Baumgartner & Steenkamp, 2001; Cheung & Rensvold, 2000; Fischer et al., 2009). The first is acquiescent responding (also called yay-saying and its opposite nay-saying) and the other is extreme responding (also called modesty responding). Acquiescent responding style (ARS) results in some up- or down-shifting of mean scores independent of item content. For example, individuals rate themselves higher (e.g., 4 on a Likert scale from 1-5) on two conflicting items such as ól am out-going in social settingsô and ól am shy when being around peopleô. Here, the expressed score is independent of their real level of shyness or extroversion. Extreme responding style (ERS) is the selection of more extreme scores (either acceptance or rejection) than would be expected based on the true level of the underlying construct. For example, a fairly extrovert individual gives a much higher rating to the item ól am out-going in social settingsô (e.g., 5 on the same scale) and much lower rating to ól am shy when being around peopleô (e.g., 1 on the same scale) than would be expected knowing his or her ótrueô level of extroversion. These tendencies can be expressions of individual differences or can be characteristics of cultural groups, indicating some cultural response or communication style (Fischer et al., 2009, Smith, 2004).

As should be obvious, there is no objective indicator of knowing the ótrueô level of a psychological construct. This fact raises many problems about interpretation and justification of standardization (to be discussed below). At the same time, the possibility that some individuals or groups are more likely to agree with items or endorse more extreme statements than they actually hold certainly creates problems in the interpretation of scores, with potentially far-reaching consequences. This is a particular concern in applied settings, like employee selection, marketing research or mental health diagnosis. For example, the wrong candidate might be selected in educational or organizational settings, or individuals might risk incorrect diagnosis on mental health checklists. Similarly, researchers are concerned because such response styles may obscure real between-group differences, or more typically will result in differences that are spurious and not related to the construct of interest. For example, it has been observed that Japanese participants typically use the middle-point category of a response scale (e.g., 3 on a Likert scale from 1-5, or 4 on a scale from 1-7), independent of item content (see Lincoln & Kalleberg, 1990). If we were to compare commitment scores of Japanese and U.S. workers, we may not find a significant difference, although behavioural indicators suggest a much higher level of commitment of Japanese workers (see Fischer & Mansell, 2009 for an overview of this literature).

A second and often paradoxical implication is that if scores of different subgroups with different response styles are combined and analyzed without considering the different score means in the subgroups, relationships might change. For example, imagine that a researcher measured the frequency of laughing and the number of times individuals cry during the week in two groups. Evidently there is a negative relationship in each group separately (the more you cry, the less you laugh). However, imagine that one group gives higher responses irrespective of item content (assuming that the overall emotional responsiveness is not different between groups). If we analyzed the data from both groups together, the relationship between crying and laughing might become zero or may even reverse due to the different positions of the two groups. Ignoring potential response styles may lead to incorrect decisions in applied settings or paradoxical findings in research. A number of standardization procedures have been proposed. Fischer (2004) provides a review of these and we will give an overview of the main types.

Score transformations typically involve an adjustment of means and/or standard deviations of either individuals or groups. Consequently, they can be grouped depending on whether they used only the means, only the standard deviations or both (see Table 1). Depending on the focus of the transformation, we can distinguish adjustments of means of either individuals, items within groups or both using either the mean across variables for each individual or across individuals within a group or both. Hence, the second important factor for classifying transformation procedures is the source of the information used for transforming scores (e.g., individuals, groups, culture; see columns in Table 1). Therefore, combining the type of statistical information used (means, standard deviations or both) with the focus of adjustment (individual, group, culture), there are a number of possibilities for adjusting raw scores. Let us consider these different possibilities briefly.

First, within-subject standardization in the first row refers to transformations of scores for each individual using the mean for that individual across all variables (Hofstede, 1980). The average across all variables for a particular individual is subtracted from the individualôs raw

score on a specific variable. The resulting adjusted score can be interpreted as the relative endorsement of this item or the relative position of the individual on a variable in relation to the other scores (Hicks, 1970). This procedure is called ipsatization (Hicks, 1970) and will yield a mean of zero across variables for this individual. If used in this way, it is supposed to control for acquiescence responding. These scores might be further adjusted for differences in the variation of the answers around the mean by dividing the ipsatized score by the standard deviation across variables for that individual (see third row, within-subject column). If the standard deviation is also controlled, the tendency to provide extreme responses is theoretically being taken care of (but see Baumgartner & Steenkamp, 2001). Box 2 shows an example of a SPSS syntax to obtain ipsatized scores.

Second, answers may be adjusted using the group mean, with the most common form being the classical z-transformation (third row, within-group column). Z-scores can be obtained by subtracting a group mean from the variable score across individuals within a group and dividing it by the standard deviation across individuals within the group (Howell, 1997). The interpretation of such adjusted scores is the relative endorsement or position of a specific individual on one variable relative to the endorsement or position of other individuals in that group. The mean of the group is zero and assuming a normal distribution of the raw responses, adjusted standard deviation will be 1. Another typical transformation is group mean centring (Aiken & West, 1991), which uses only the mean. SPSS produces z-scores automaticaly through the descriptive option (click on ósave standardized values as variablesô under óAnalysisô, óDescriptive statisticsô, óDescriptivesô).

Table 1. *Score transformation procedures*

Statistical information used for adjustment	Within-subject (adjustment across variables for each individual)	Within-group (adjustment across individuals for one variable)	Within-culture (adjustment across individuals and variables)	Double standardization (Leung & Bond, 1989)
Mean	yô = x-mean$_{individual}$ *Ipsatization*	yô = x-mean$_{group}$ *Group mean centring*	yô = x-mean$_{culture}$ *Grand mean centring*	yô = x-mean$_{individual}$ yôô = yô-mean$_{yô\,culture}$
Dispersion indices (commonly standard deviation)	yô = x/dispersion$_{individual}$	yô = x/dispersion$_{group}$	yô = x/dispersion$_{culture}$	yô = x/dispersion$_{individual}$ yôô = yô/dispersion$_{yô\,culture}$
Means and dispersion indices	yô = (x-mean$_{individual}$)/dispersion $_{individual}$ *Ipsatisation*	yô = (x-mean $_{group}$)/dispersion $_{group}$ *Z-transformation*	yô = (x-mean-$_{culture}$)/dispersion $_{culture}$	yô = (x-mean$_{group}$)/ dispersion $_{group}$ yôô = (x-mean $_{yô\,group}$) / dispersion $_{yô\,group}$

Third, within-culture standardization (Bond, 1988; Leung & Bond, 1989) is similar to centring and z-transformations, but uses the mean across all items and individuals within a group compared with the mean across individuals on one variable or item only (as done in z-transformation). As before, transformations can use only the mean (first row), only the grand standard deviation (second row) or both (third row). Table 1 provides the formula to be input in a variant of the syntax in Box 2 (the overall mean needs to be computed separately and input instead of the ómeanô command).

Finally, Leung and Bond (1989) introduced double standardization which is a combination of within-subject and within-culture standardization. First, raw scores are transformed within the individual (within-subject standardization) and these scores are then adjusted using the group mean across individuals and variables (within-culture standardization). This will yield a mean of zero for each individual across variables as well as for each variable across individuals. Assuming normality of the raw data, the standard deviations should be 1 for both individuals across variables and variables across individuals. Syntax in Box 2 can be used to achieve this transformation (note that the procedure now involves two steps, first adjustment of individual, then group means across all items and individuals).

Box 2: Creating ipsatized scores

Example of SPSS syntax (to be copied and pasted into a SPSS syntax window):

Controlling for mean only (row 2 and column 2 in Table 1)

Compute *Name=variable_1*-(mean(*item_1*,*item_2*, é , *item_x*)).

Controlling for standard deviation (row 3 and column 2 in Table 1)

Compute *Name=variable_1*/SD(*item_1*,*item_2*, é , *item_x*).

Controlling for mean and standard deviation (row 4 and column 2 in Table 1)

Compute *Name=variable_1*-(mean(*item_1*,*item_2*, é , *item_x*)/SD(*item_1*,*item_2*, é , *item_x*)).

For *name* type the name your ipsatized variable should take. *Variable_1* is the raw score variable (either individual item or sum/average of a number of items) that is being standardized. The items in the parentheses are the items that are used to standardize item 1.

In summary, there are three main groups of score transformation procedures based on whether they use (1) means, (2) standard deviations, or (3) both means and standard deviations. Depending on the focus, we can also distinguish within-subject (adjustment across variables for each individual), within-group (adjustment across individuals for each variable) and within-culture (adjustments across individuals and variables within a culture). Finally, double standardization combines both within-subject and within-culture adjustment.

A final but less frequently recommended and used method is to use covariate analyses to adjust for the overall response tendency of the individual (Fischer, 2004). Similar to within-subject standardization, the overall mean across all items or scores for an individual is created and then partialled out in an analysis of covariance or partial correlation. This can be done fairly easily in standard statistical programmes such as SPSS.

The use of these techniques has been increasing over the last three decades, even when accounting for the increased publication rate (Fischer, 2004). It certainly reflects the increased trend to use self-reports in psychological and social research, but also demonstrates an increasing awareness among researchers about the problems of using raw scores to make cross-cultural or cross-ethnic comparisons. For example, the most common types of transformations used in studies published in the Journal of Cross-Cultural Psychology are ipsatization, followed by double standardization and grand mean/standard deviation centring (Fischer, 2004). Ipsatization is also common in personality research (e.g., Bartram, 1996). The rationale for using either means or standard deviations when adjusting raw data is different. Adjustment using means is typically used if researchers are concerned with acquiescent bias. If the mean across all items for one cultural group is consistently higher or lower compared with the means from another group, adjusting for these mean differences might be advocated (Hofstede, 1980). Z-transformation, or grand-mean centring, can be used if there are concerns with ARS.

The rationale for using standard deviations is to adjust for extreme response bias (Kashima et al., 1992), however such transformation using only the standard deviation is very rare and it is hard to detect ERS. The most straightforward method for detecting ERS would be a relative balance of positively and negatively phrased items in the instrument used (Chun et al., 1974). Observing an overall difference in standard deviations would make extreme response tendencies a possible explanation. Using only items phrased in one direction (either only positively or only negatively) makes detection of extreme responding more difficult, especially considering ARS and ERS often covary (Smith, 2004). In such a case, adjustment using both means and standard deviations might be appropriate. Fischer (2004) discusses these issues in more detail.

Nevertheless, the use of these score adjustment procedures is not without problems and researchers should exercise caution when using them. In the next section we discuss some implications of such transformations for using them in further statistical tests. We distinguish here between structure-oriented tests (such as exploratory [EFA] or confirmatory factor analysis [CFA], multidimensional scaling [MDS]) and level-oriented tests (such as ANOVA and regression). We structure this discussion by the type of transformation (combining both within-subject & double-standardization, and within-group & within-culture).

IMPLICATIONS OF WITHIN-SUBJECT STANDARDIZATION AND DOUBLE STANDARDIZATION

Within-subject standardization yields ipsative scores (Hicks, 1970). These scores are problematic in a number of ways. Hicks (1970) was the first to highlight these issues in some more detail, but these issues have received much attention subsequently (see for example, Baron, 1996; Bartram, 1996; Chan, 2003; Closs, 1996; Cornwell & Dunlop, 1994; Tenopyr, 1988).

When using ipsatisation, the mean for each individual will be zero and ñeach score for an individual is dependent on his own scores on other variables, but is independent of, and not comparable with, the scores of other individualsò (Hicks, 1970, p. 167). This leads to the sum of variances and covariances being zero in every row and column of the covariance matrix, resulting in a singular matrix for which no regular inverse can be computed. This makes the resulting matrix unsuitable for factor analysis (Chan, 2003). A least one of the *k*-1 (with *k* being the number of variables) covariance terms in each row and column will be negative, independent of substantive relationships. This implies that at least one covariance (or correlation) is due to methodological artifacts caused by the ipsatisation procedure rather than the true relationship between constructs. The average item-intercorrelation is predictable knowing the number of variables standardised: average $r = -1/(k$ ï $1)$; where *k* is the number of variables, assuming equal variances (Hicks, 1970). This creates problems when such matrices are used with correlational

techniques such as EFA and CFA as well as multivariate techniques such multiple regression and MANOVA (Closs, 1996; Cornwell & Dunlap, 1994).

An interpretational issue is that the overall score for that person across all items is zero. Using the crying versus laughing example again, standardizing the responses of individuals, the adjusted score after ipsatization could be interpreted as the relative frequency of laughing compared to crying (and vice versa). The score then expresses the frequency of crying in relation to the frequency of laughing. In addition to the noted statistical problems, this also has important theoretical implications. Applying within-subject standardization or ipsatization, a researcher assumes a limited resource, or fixed-size pie scenario. This may or may not make sense in particular areas of investigation. Using the crying-laughing example, we would assume a) that all people have the same level of emotionality and b) the relative score then also implies that people spend all their time either crying or laughing. In motivational research it may be logical to assume that individuals have only a limited area of energy and therefore there is a balance or trade-off for where this energy can be invested. However, a different motivation researcher might assume that there are individuals that have higher levels of motivation overall and would like to predict what personality variables are associated with this overall level of motivation. In this case, ipsatized scores would not be useful.

Another problem is that if all the scores are influenced by an unmeasured third variable, it may also obscure any differences and lead to paradoxical results after standardization. Standardizing the length of body features of a mouse and an elephant might lead to the conclusion that the tail of the mouse is longer than that of the elephant. However, since both are related to the overall size of the animal, this statement is certainly incorrect (if scores are interpreted as absolute measures). This form of standardization therefore raises important theoretical questions about the meaning and interpretation of such scores. Note the implications when comparing scores across groups (such as when using t-tests or ANOVA) since mean scores can be severely over- or underestimated.

There are also issues related to double standardization. Of particular concern is that double standardized scores have been recommended in the cross-cultural literature for detecting so-called óculture freeô (Hasegawa & Gudykunst, 1998; Leung & Bond, 1989; Singelis, Bond, Sharkey & Lai, 1999; Triandis et al., 1993) or óeticô dimensions (Yamaguchi, Kuhlman & Sugimori, 1995). This argument has to be seriously questioned. Fischer (2004) demonstrated that the structure of the horizontal-vertical individual-collectivism scale (Triandis & Gelfand, 1998) can not be adequately recovered when using double standardization. To demonstrate this point further, using data from 150 undergraduate students who completed a version of the Big Five Inventory (BFI, Goldberg, 1999), we tested whether we could recover the initial factor structure in the same data set after within-subject standardization. The questionnaire consists of 50 items, half of which are negatively phrased and the items across the five domains have an average inter-item intercorrelation of .02. According to Bartramôs (1996) analysis, this data set might be suitable for ipsatization. We extracted five factors from the ipsatized data matrix and used Tuckerôs Phi to rotate the structure of the ipsatized matrix to the raw matrix (for procedures on how to do this, see below). Values of .95 or higher are normally seen as indicators of good factorial agreement (Van de Vijver & Poortinga, 2002). The values obtained were .97; .77; .92; .98 and .86. Therefore, three of the five factors were not adequately recovered. Hence, so-called óculture freeô correlation matrices based on within-subject and doubly standardization (ipsatization) yield different factor structures even within the same data set and resulting factor structures are highly ambiguous (cf. Closs, 1996; Cornwell & Dunlop, 1994), even with data sets that might be suitable according to some recommendations (e.g., Bartram, 1996).

Turning to CFA, Chan (2003) proposed a method which allows the use of ipsative data matrices. The method involves a number of constraints to be placed on the factor loading and error covariance matrix. Chan reports an adequate recovery of the initial factor structure, however, the fit indices differ considerably between the raw and standardized matrix solution. Further research is thus needed to investigate this promising possibility further.

If researchers are interested in investigating the underlying structure of ipsative data matrices, non-parametric techniques such as multi-dimensional scaling could be used. Multi-dimensional scaling (Borg & Groenen, 2005; Kruskal & Wish, 1978), using distances and specification of ordinal measurement, is an appropriate alternative for analysing ipsative data. Furthermore, multi-dimensional scaling can be used to assess structural relations among variables instead of factor analysis if researchers are concerned that response biases might obscure structural relationships in the data because this technique is not influenced by overall score level differences in different groups. Therefore, MDS can be used with ipsative matrices, but EFA and CFA should not be used with ipsatised matrices because this produces spurious results that in most cases do not correspond to substantial relationships (see Closs, 1996; Cornwell & Dunlop, 1994).

Implications of within-group and within-culture standardization

Although these standardization procedures are less frequently used (as judged by publications in the Journal of Cross-Cultural Psychology, Fischer, 2004), they do not have the same problematic statistical properties as ipsative scores. Centering can be used to remove patterning effects (e.g., Leung & Bond, 1989). For example, in the above example of laughing and crying, we could centre the scores in each group separately and thereby remove the mean differences. EFA, CFA and MDS can all deal with data that has been centered. Z-transformation is particularly interesting. In the example noted above, we could just transform the scores to z-scores in each sample separately and then conduct a factor analysis in the combined sample. This works quite effectively to eliminate ARS (Groenvynck, & Fontaine, 2003). MDS is less affected by ARS in the first place and therefore can be used without further problems.

SUMMARY

As discussed, standardization may address some of the problems caused by different response styles. However, we also noted that it is hard to establish whether scores are being affected by response styles. For research purposes, it is better to model such response style factors and estimate the effect directly (e.g., Welkenhuysen-Gybels, Billiet and Cambre, 2003) rather than controlling for it without knowing the exact extent and nature of it. It is also possible to use within-subject designs to control for individual differences in reaction to stimuli (see Howell, 1999).

CONCLUSIONS

Standardization has multiple meanings and each form has its place in research methods. The purpose of the current article was to clarify the usage of the term and highlight some of the applications and the impact that especially standardization of scores has on multivariate techniques.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage Publications.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology, 69,* 49-56.

Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational and Organizational Psychology, 69*, 25-39.

Baumgartner, H., & Steenkamp, J. B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 37*, 143-156.

Bond, M. H. (1988). Finding universal dimensions of individual variation in multicultural studies of values: The Rokeach and Chinese values surveys. *Journal of Personality and Social Psychology, 55*, 1009-1015.

Borg, I., & Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications* (2nd ed.). New York: Springer.

Brouwers, S. A., van de Vijver, F. J. R., & van Hemert, D. A. (2009). Variation in Raven's progressive matrices scores across time and place. *Learning and Individual Differences, 19*, 330-338.

Chan, W. (2003). Analyzing ipsative data in psychological research. *Behaviormetrika, 30*, 99-121.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31*, 160-186.

Chun, K-T., Campbell, J. B., & Yoo, J. H. (1974). Extreme response style in cross-cultural research: A reminder. *Journal of Cross-Cultural Psychology, 5*, 465-479.

Closs, S. J. (1996). On the factoring and interpretation of ipsative data. *Journal of Occupational and Organizational Psychology, 69*, 41-47.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problems of units and the circumstance for POMP. *Multivariate Behavioral Research, 34*, 315-346.

Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville and Willson (1991). *Journal of Occupational and Organizational Psychology, 67*, 89-100.

Fischer, R. & Mansell, A. (2009). Commitment across cultures: A meta-analytical approach. *Journal of International Business Studies, 40*, 1339-1358.

Fischer, R. (2004). Standardization to account for cross-cultural response bias: A classification of score adjustment procedures and review of research in JCCP. *Journal of Cross-Cultural Psychology, 35*, 263-282.

Fischer, R. & Fontaine, J. J. R. (2010). Methods for investigating structural equivalence. In D. Matsumoto & F. van de Vijver (eds), Handbook of Cross-Cultural Research Methods. Oxford University Press.

Fischer, R., Fontaine, J., van de Vijver, F. J. R., & van Hemert, D. (2009). What is Style and What is Bias in Cross-Cultural Comparisons? An Examination of Response Styles in Cross-Cultural Research. In A. Gari & K. Mylonas (eds), *Quod Erat Demonstrandum: From Herodotusô Ethnographic Journeys to Cross-Cultural Research* (pp.137-148). Athens: Pedio.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I.

Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe, Vol. 7* (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Groenvynck, H., & Fontaine, J. R. J. (2003, July). *Can centrering eliminate acquiescence in cross-cultural research?* Paper presented at the 6th Regional Congress of the International Association for Cross-Cultural Psychology, Budapest, Hungary.

Harris, P. R. (In Press). *Designing and Reporting Experiments in Psychology*. Third edition: Milton Keynes: Open University Press.

Hasegawa, T., & Gudykunst, W. B. (1998). Silence in Japan and the United States. *Journal of Cross-Cultural Psychology*, 29, 668-684.

Hicks, L. E. (1970). Some properties of ipsative, normative and forced-choice normative measures. *Psychological Bulletin*, *74*, 167-184.

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

Hofstee, W. K. B. (2008). *Restyling personality assessments* (Chapter 9, pp. 223-235). In L. B. Palcroft & M. V. Lopez (Eds.). Personality Assessment: New Research. New York: Nova Science Publishers.

Howell, D.C. (1997). *Statistical methods for psychology*. Belmont, CA: Duxbury Press.

Kashima, Y., Siegal, M., Tanaka, K., & Kashima, E. S. (1992). Do people believe behaviors are consitent with attitudes? Towards a cultural psychology of attribution processes. *British Journal of Social Psychology, 31*, 111-124.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Sage University Paper No. 07. London: Sage.

Lamiell, J. T. (2007). On sustaining critical discourse with mainstream personality investigators. *Theory and Psychology*, *17*, 2, 169-185.

Leung, K. & Bond, M. H. (1989). On the empirical identification of dimensions for cross-cultural comparisons. *Journal of Cross-cultural psychology, 20*, 133-151.

Lincoln, J. R., & Kalleberg, A. L. (1985). Work organization and workforce commitment: A study of plants and employees in the US and Japan. *American Sociological Review, 50*, 738-760.

Murphy, K. R. & Davidshofer, C. O. (1994). *Psychological testing: Principles and applications*. Englewood Cliffs, NJ: Prentice-Hall.

Singelis, T. M., Bond, M. H., Sharkey, W. F., & Lai, C. S. Y. (1999). Unpackaging cultures's influence on self-esteem and embarrassability. *Journal of Cross-Cultural Psychology, 30*, 315-341.

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology, 35*, 50-61.

Tenopyr, M. L. (1988). Artefactual reliability of forced-choice scales. *Journal of Applied Psychology, 73*, 749-751.

Triandis, H. C., & Gelfand, M. J. (1998). Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology, 74*, 118-128.

Welkenhuysen-Gybels, J., Billiet, J., & Cambre, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-Cultural Psychology, 34*, 702-722.

Yamaguchi, S., Kuhlman, D. M., & Sugimori, S. (1995). Personality correlates of allocentric tendencies in individualist and collectivist cultures. *Journal of Cross-Cultural Psychology, 26*, 658-672.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology, 33*, 141-156.

Heathcote, A. (1996). RTSYS: A DOS application for the analysis of reaction time data. *Behavior Research Methods, Instruments, and Computers*, *28*, 427-445.

Heeren, T. & D'Agostino R. (1987). Robustness of the two independent samples t-test when applied to ordinal scaled data. *Statistics in Medicine*, *6*, 79-90.

Olivier, J., Johnson, W.D., Marshall, G.D. (2008). The logarithmic transformation and the geometric mean in reporting experimental IgE results: What are they and when and why to use them? *Annals of Allergy, Asthma and Immunology*, *100*, 333-337. Erratum in: *Annals of Allergy, Asthma and Immunology*, 100, 625-626.

SAS Institute Inc. (2008). *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Wackerly, D., Mendenhall W., and Scheaffer R. (2007). *Mathematical Statistics with Applications, 7th Edition*. Belmont, CA: Duxbury.