Int J Public Health (2013) 58:637–642
DOI 10.1007/s00038-012-0416-3

HINTS & KINKS

The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown?

Rob Eisinga · Manfred te Grotenhuis · Ben Pelzer

Received: 10 January 2012 / Revised: 27 June 2012 / Accepted: 24 September 2012 / Published online: 23 October 2012

Introduction

To obtain reliable measures researchers prefer multiple-item questionnaires rather than single-item tests. Multiple-item questionnaires may be costly however and time-consuming for participants to complete. They therefore frequently administer two-item measures, the reliability of which is commonly assessed by computing a reliability coefficient. There is some disagreement, however, what the most appropriate indicator of scale reliability is when a measure is composed of two items. The most frequently reported reliability statistic for multiple-item scales is Cronbach's coefficient alpha and many researchers report this coefficient for their two-item measure (Cuijpers et al. 2009; Löwe et al. 2005; Michal et al. 2010; Young et al. 2009). Others however claim that coefficient alpha is inappropriate and meaningless for two-item scales (Sainfort and Booske 2000; Verhoef 2003; Cramer et al. 2006; O'Brien et al. 2008). Instead, they recommend using the Pearson correlation coefficient as a measure of reliability. Still others argue that the inter-item correlation equals the split-half reliability estimate for the two-item measure and they advocate the use of the Spearman-Brown formula to estimate the reliability of the total scale (Hulin et al. 2001). As these recommendations are reported without elaborating, there is considerable confusion among end users as to the best reliability coefficient for two-item measures. This note aims to clarify the issue.

It is important to emphasize at the outset that it is not our intention in this article to endorse or to promote the use of two-item scales. Quite the contrary, having only two items to identify an underlying construct has been recognized as problematic for some time and we support the claim that using more items is better, particularly in exploratory research (Herbert et al. 1998; Little et al. 1999; Emons et al. 2007). The use of multiple, heterogeneous indicators enhances construct validity in the sense that it increases the likelihood of adequately identifying the construct of interest. Also, assessments used for individual diagnosis, tracking or admission purposes and high-stakes decision-making require ample information about the individual and this necessarily implies the application of long tests or inventories (Emons et al. 2007). However, in large-scale health surveys for example, resource and survey time constraints often mean that only a limited number of items can be included to assess a particular construct and it is not at all uncommon to find questionnaires having no more than two indicators to gauge a particular self-assessment. Further, it is a common situation facing researchers that poor quality items have to be removed from a limited item pool, resulting in scales with a small number of items, occasionally two. Our concern is how to best estimate reliability in this actual practice setting. We assume in our discussion that the available data are such that it is justified to calculate a reliability estimate. Hence, we ignore empirical issues such as nonlinear relationships, notoriously non-normal distributions, small sample sizes and like complications that prohibit meaningful reliability calculation and inference.

For a reliability coefficient to accurately reflect the true reliability of a two-item scale, the observations have to meet particular requirements. Classical test theory summarizes these requirements in measurement models (Lord and Novick 1968). We briefly discuss these models and subsequently present data examples that meet their assumptions.

R. Eisinga (✉) · M. t. Grotenhuis · B. Pelzer
Radboud University Nijmegen, P.O. Box 9104,
6500 HE Nijmegen, Netherlands
e-mail: r.eisinga@maw.ru.nl

This procedure allows us to evaluate the appropriateness of the reliability estimates for two-item scales. The results we report should be useful to researchers, not in the least because the issue frequently turns up in reviewers' comments to submitted journal papers (Hulin et al. 2001).

Measures

According to classical test theory, the observed score (y) on an item is equal to the sum of a true score (τ) and a measurement error (ε). If the measure is unbiased, the expected value of the error is zero (i.e., $E(\varepsilon) = 0$). If we have a summated two-item scale and y_i is the observed score on item i and Y is the scale score, then

$$Y = y_1 + y_2 = (\tau_1 + \varepsilon_1) + (\tau_2 + \varepsilon_2),$$

where it is assumed that $\mathrm{Cov}(\varepsilon_1, \varepsilon_2) = 0$, meaning that the errors are independent across items, and that $\mathrm{Cov}(\tau, \varepsilon_i) = 0$, thus true score and errors are also uncorrelated. If τ_1 and τ_2 are measures of the same underlying true score, then the only difference between the two items is a matter of scaling or item difficulty. Hence, we can think of a single true score τ that is the same for the two items but where τ is multiplied by different constants λ_i for item 1 and item 2, or where different constants s_i are being added to τ. We therefore have

$$\tau_1 = \lambda_1\tau + s_1 \quad \text{and} \quad \tau_2 = \lambda_2\tau + s_2.$$

Such transformations to the true score obviously result in τ_1 being unequal to τ_2 even though they are measures of the same true score τ, which is imperfectly measured only as a result of measurement error. Together, true score and measurement error, possibly subject to some transformation, constitute a measurement model. The major ones in test theory include parallel, (essentially) tau-equivalent, and congeneric measures (Lord and Novick 1968).

The measures comprising a two-item scale are strictly *parallel* if $\tau_1 = \tau_2$ and $\mathrm{Var}(\varepsilon_1) = \mathrm{Var}(\varepsilon_2)$. These conditions imply that the amount of variation in the observed item score that is determined by the true score is the same for the two items and, additionally, that the expected values of the two items are equivalent. The assumption of *tau-equivalence* also implies that each person has a constant true score over items, but the measurement error variances may vary across items, i.e., $\mathrm{Var}(\varepsilon_1) \neq \mathrm{Var}(\varepsilon_2)$. *Essentially tau-equivalence* holds if each person's true score for item 1 differs by an additive constant from the true score for item 2 (i.e., $s_1 \neq s_2$). It implies that whereas the true scores differ across items, true-score variance is constant. The error variances however differ. Finally, *congeneric* measures assume that for each person the true score may vary across items but there is an additive and a multiplicative constant that relates the true scores across any two items. Neither true-score nor error variances need to be equal. Hence, the congeneric case implies that $\lambda_1 \neq \lambda_2$ and that $\mathrm{Var}(\varepsilon_1) \neq \mathrm{Var}(\varepsilon_2)$.

Reliability estimates

To evaluate the implications for reliability, we present an example for each of the measurement models. The observed score for each of the two items y_1 and y_2 is the sum of a true score, possibly subject to some linear transformation (λ_i), and an error term, possibly multiplied by some factor (λ_i^{ε}), but with an expected value of zero. The scale score Y is equal to their unweighted sum. In our example of parallel measures, we assume that the observed item score is $\lambda_i = 0.8$ times the true score, with $\mathrm{Var}(\tau_i) = 1$, and we multiplied the error by $\lambda_i^{\varepsilon} = \sqrt{1 - 0.8^2} = 0.6$ This still implies that $\tau_1 = \tau_2$ and that $\mathrm{Var}(\varepsilon_1) = \mathrm{Var}(\varepsilon_2)$. The summary statistics and the reliability calculations are presented in Table 1.

As can be seen in the top part of the table, the means and the variances of the parallel items are the same. If we square the correlation between the true score and the scale score ($r_{\tau Y} = 0.883$, so $r^2_{\tau Y} = 0.780$), we obtain the true reliability that is identical to the calculation of Cronbach's coefficient alpha ($\alpha_{y1y2} = 0.780$). This finding is consistent with the definition of reliability as the proportion of the variance in the observed scale score that is explained by variation in the true score. The Pearson correlation between y_1 and y_2 ($r_{y1y2} = 0.640$) is seen to be lower than the reliability of the two-item scale. The coefficient equals the squared-correlation between the true score and a single-item score and it thus represents the amount of variation in the single item that is determined by the true score. Hence, the Pearson correlation is not an adequate measure of the reliability of a two-item scale. Rather, one could consider it to be the reliability of a one-item test.

If two items are parallel, the inter-item correlation represents the correlation between one half of the test with the other half, i.e., the split-half reliability of the scale (Hulin et al. 2001). Given this correlation, we may easily convert a split-half reliability into a reliability that has the coefficient alpha interpretation using the Spearman-Brown formula, given in Table 1 ($\rho_{y1y2} = 0.780$). For two-item scales this estimate is equivalent to standardized coefficient alpha based on standardized items. It is not true however, as some authors have suggested, that for two-item scales the Spearman-Brown coefficient is the equivalent of coefficient alpha (Rüsch et al. 2009). This is only true if $\mathrm{Var}(\varepsilon_1) = \mathrm{Var}(\varepsilon_2)$, as is the case if the items are parallel in the true-score sense of parallel measures.

When measures are tau-equivalent, then $\tau_1 = \tau_2$ but the assumption that $\mathrm{Var}(\varepsilon_1) = \mathrm{Var}(\varepsilon_2)$ is relaxed. To make the

Table 1 Measures and calculation of reliability for two-item scale

Measure	Item scores	$E(y_i)$	$\text{Var}(y_i)$	$\text{Cov}(y_1, y_2)$	Pearson r_{y1y2}	Cronbach α_{y1y2}	Spearman-Brown ρ_{y1y2}	True reliability $r^2_{\tau Y}$
Parallel	$y_1 = 0.8\tau + 0.6\varepsilon_1$	0	1	0.640	0.640	0.780	0.780	0.780
	$y_2 = 0.8\tau + 0.6\varepsilon_2$	0	1					
Tau-equivalent	$y_1 = 0.8\tau + 0.6\varepsilon_1$	0	1	0.640	0.749	0.850	0.857	0.850
	$y_2 = 0.8\tau + 0.3\varepsilon_2$	0	0.730					
Essentially	$y_1 = 0.8\tau + 0.6\varepsilon_1 + 1$	1	1	0.640	0.749	0.850	0.857	0.850
Tau-equivalent	$y_2 = 0.8\tau + 0.3\varepsilon_2$	0	0.730					
Congeneric	$y_1 = 0.8\tau + 0.6\varepsilon_1$	0	1	0.160	0.444	0.441	0.615	0.690
	$y_2 = 0.2\tau + 0.3\varepsilon_2$	0	0.130					
	$y_1 = 0.8\tau + 0.6\varepsilon_1$	0	1	0.480	0.716	0.797	0.834	0.813
	$y_2 = 0.6\tau + 0.3\varepsilon_2$	0	0.450					
	$Y = y_1 + y_2$	$r_{y1y2} = \frac{\text{Cov}(y_1y_2)}{\sqrt{\text{Var}_{(y_1)} \times \text{Var}_{(y_2)}}}$		$\alpha_{y1y2} = \frac{4\,\text{Cov}(y_1y_2)}{\text{Var}_{(y_1)} + \text{Var}_{(y_2)} + 2\,\text{Cov}_{(y_1y_2)}} \rho_{y1y2} = \frac{2\,r_{y1y2}}{1 + r_{y1y2}}$				
	Substituting r_{y1y2} into ρ_{y1y2} gives							
	$\rho_{y1y2} = \frac{4\,\text{Cov}(y_1y_2)}{2\sqrt{\text{Var}(y_1) \times \text{Var}(y_2)} + 2\,\text{Cov}(y_1y_2)}$ As $\frac{\text{Var}(y_1) + \text{Var}(y_2)}{2} \geq \sqrt{\text{Var}(y_1) \times \text{Var}(y_2)} \rightarrow \rho_{y1y2} \geq \alpha_{y1y2}$ (arithmetic mean) (geometric mean)							

error variances differ, the error terms were multiplied by different constants. As can be seen in Table 1, the variances of the items are no longer identical. However, the squared-correlation between the true score and the scale score ($r^2_{\tau Y} = 0.850$) again equals coefficient alpha ($\alpha_{y1y2} = 0.850$). Similar results go for essentially tau-equivalent measures. The inclusion of an additive constant affects the item means, but it is irrelevant for their variances and covariances. As reliability is a variance-accounted-for statistic, it is unaffected by unequal additive constants.

Coefficient alpha is an estimate of the reliability of a sum of parallel or (essentially) tau-equivalent measures (Bollen 1989). Hence it assumes that the two items measure the same construct on the same scale, with the only variance unique to an item being completely comprised of measurement error. The implication of this restrictive assumption may be gauged by examining the results for congeneric measures, that relax both the assumption that $\tau_1 = \tau_2$ and that $\text{Var}(\varepsilon_1) = \text{Var}(\varepsilon_2)$. Table 1 presents two examples. The results were obtained by multiplying both the true score and the error terms by different constants.

The first example shows that for congeneric measures coefficient alpha ($\alpha_{y1y2} = 0.441$) may be substantially smaller than the squared-correlation between the true score and the scale score ($r^2_{\tau Y} = 0.690$). That is, coefficient alpha is a lower-bound estimate that always underestimates the true reliability of a scale when measures are congeneric (Bollen 1989; Sijtsma 2009; Revelle and Zinbarg 2009). For a two-item scale the Spearman-Brown coefficient is always larger than coefficient alpha (See Table 1), except for the case when $\text{Var}(\varepsilon_1) = \text{Var}(\varepsilon_2)$. The Spearman-Brown formula assumes that the split-halves are parallel measures. If this assumption is violated the formula does not hold and the coefficient may either underestimate (Table 1: congeneric example 1) or overestimate (example 2) the true reliability of the composite scale.

The difference between the true reliability and the estimate obtained by coefficient alpha or the Spearman-Brown formula is the coefficient's bias. To examine the bias of the reliability statistics for both tau-equivalent and congeneric measures, we multiplied the true score and the error terms by 1.6×10^9 different values for λ_i and λ_i^{ε}. The values were obtained by generating all possible combinations of λ_1, λ_2, λ_1^{ε} and λ_2^{ε}, each of which is equidistantly spaced in the interval [0,1], a distance .005 apart. For tau-equivalent measures $\lambda_1 = \lambda_2$.

Figure 1 displays the relationships between the mean and the standard deviation of the bias and the two-item Pearson correlation. The graph and the bias formula below the graph indicate that coefficient alpha is unbiased when measures are at least tau-equivalent, hence if $\lambda_1 = \lambda_2$. The Spearman-Brown coefficient is found on average to slightly overestimate reliability if the two-item scale has tau-equivalent items. The same figure also shows that if items are congeneric, coefficient alpha tends to have a much larger bias than the Spearman-Brown statistic. Also, whereas the Spearman-Brown coefficient becomes progressively more precise and, by and large, more unbiased as the correlation between the two congeneric items increases, the underestimation of coefficient alpha remains substantial even if the inter-item relationship is rather strong.

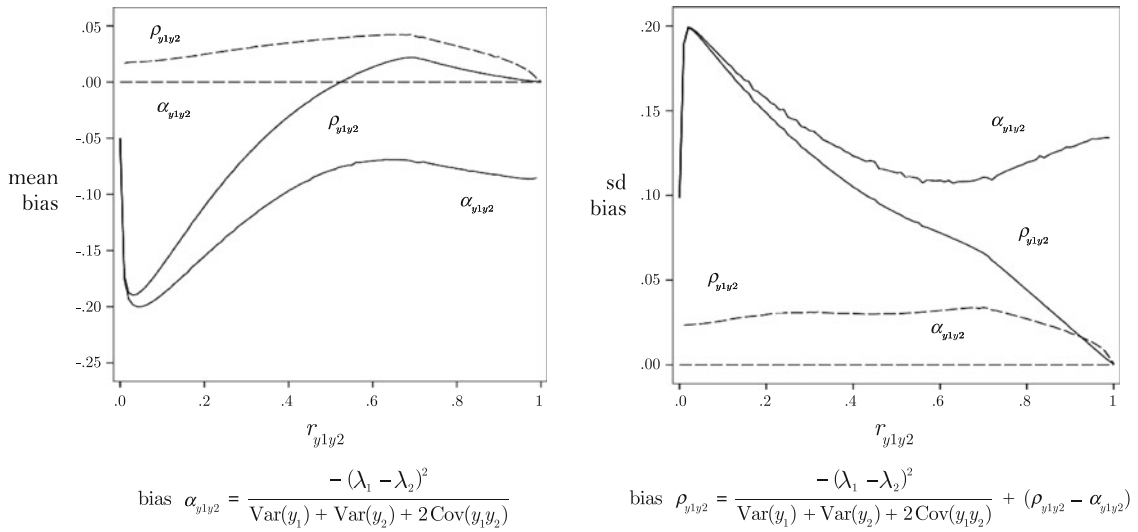

Fig. 1 Mean and standard deviation of the bias of Cronbach's coefficient alpha and the Spearman-Brown coefficient by Pearson correlation for tau-equivalent (*dashed lines*) and congeneric (*solid lines*) items

Hence, we have the seemingly contradictory result that the coefficient with the strongest assumptions performs better than the coefficient with more relaxed assumptions if the assumptions in question are violated. This apparent paradox is reconciled by the observation that coefficient alpha is a lower bound of true reliability and that, in the two-item case, the Spearman-Brown estimate is always greater than or equal to alpha. The underestimation by coefficient alpha is, on average, larger than the misestimation by the Spearman-Brown statistic. We may therefore conclude that, as the conditions of essentially tau-equivalence typically fail to fit actual data encountered in practice, the Spearman-Brown formula is a more appropriate reliability coefficient to report for a two-item scale.

Given the inter-item Pearson correlation the Spearman-Brown reliability coefficient is easy to calculate by hand using $\rho_{y1y2} = 2r_{y1y2}/(1 + r_{y1y2})$. For two-item scales, the Spearman-Brown statistic may also be expressed as

$$\rho_{y1y2} = 1/\left[1 + \left(1/\left[\frac{r_{y1y2}}{(1 - r_{y1y2})} + \frac{r_{y1y2}}{(1 - r_{y1y2})}\right]\right)\right],$$

where the term $r_{y1y2}/(1 - r_{y1y2})$ represents a ratio of the proportion of the variance in a single item explained by the true score (i.e., the individual item's reliability) to the proportion unexplained, turning the Spearman-Brown coefficient into an aggregate measure of such information. This representation of ρ_{y1y2} may be recognized as being equal to Hancock-Mueller's reliability coefficient H, if the factor loadings of the two items are constrained to be equal (i.e., tau-equivalence constraint), implying that the squared standardized factor loadings equal the Pearson correlation (Hancock and Mueller 2001). Under the assumption of a tau-equivalent pair of two items, the largest eigenvalue is simply $1 + r_{y1y2}$ and the item's variance explained by the common factor thus equals $(1 + r_{y1y2})/2$. It is important mentioning in this context that without equality constraint the underlying construct is not properly identified in factor analysis such that a unique factor solution cannot be recovered. Constraining the loadings of the two items to be equal is justified only if the assumption of tau-equivalence is satisfied. Unfortunately, there is no way to test this assumption with only two items, as there are too few observed covariances.

What is equally stringent for a two-item scale is the classical test theory assumption that the items are locally independent. The principle of local independence means that there should not be any correlation between two items after the effect of the underlying construct is partialled out, i.e., the correlation of residuals should be zero (Lord and Novick 1968; Embretson and Reise 2000). In other words, the items should only be correlated through the construct the scale or test is measuring. If there are significant correlations among the items after the contribution of the common latent variable is removed, i.e., among the residuals, then there is a specific factor or subsidiary construct in the measurement which is not accounted for by the common factor. An example of local dependence arises when two items have highly similar item wordings. Participants may respond to the second item in the same way as to the first item without regard to the common underlying construct. That is, their responses are linked for reasons beyond a common construct and influenced by a specific factor having little to do with the latent factor of interest. Local dependence must be guarded against because its occurrence leads to a false impression of a scale's psychometric properties. Specifically, dependency

among items inflates reliability estimates and it may thus give a fake impression of the precision and quality of the scale. Researchers should therefore attempt to limit the occurrence of local dependence by formulating items whose responses depend only on the participant's position on the latent factor and not on the response to another item.

We know of no statistical procedure for detecting violation of the local independence assumption if the scale has only two items. Violation arises primarily from two items that share variance even after extracting a common factor. For a pair of two items, however, one single factor completely accounts for the inter-item covariance. An important issue to consider here is whether this factor is a valid representation of the construct one is trying to identify. Two items are necessarily statistically independent once the common factor has been extracted from the observed covariance. This does not imply that the items are locally independent however. It only means that is not possible to test this assumption for a scale with two items. This is yet another issue that argues against the use of two-item scales.

Finally, the relationship between bias and the Pearson correlation visualized in Fig. 1 should not be taken to mean that it is desirable to use items with as strong as possible correlation. An increase in correlation between two items may be accompanied by a decrease in content validity, i.e., the extent to which a concept is represented by the items. Items should be univocal, that is, measure one and only one thing that completely accounts for their covariation, and as heterogeneous as possible within the limits of the definition of what one is trying to measure rather than maximum homogeneous in the statistical sense.

Conclusion

The Pearson correlation is not an adequate measure of the reliability of a two-item scale. Rather, one may call that the reliability of a one-item test. Cronbach's alpha is an accurate estimate of reliability under rather restrictive assumptions. As these conditions are typically too much to expect from a composite scale, coefficient alpha almost always underestimates true reliability, sometimes rather substantially (Bollen 1989; Sijtsma 2009; Revelle and Zinbarg 2009). Obviously, the same goes for statistics that are the equivalent of coefficient alpha for two-item scales such as Guttman's lambda-2. Although they are often close in size, for two-item measures the Spearman-Brown coefficient is never lower than coefficient alpha and almost always higher. It is also, on average, less biased, especially if the correlation between the items is relatively strong. Hence, the most appropriate reliability statistic for a two-item scale is the Spearman-Brown coefficient that together with standardized coefficient alpha, its equivalent for two-item measures, is offered by software such as *SPSS*, *SAS*, and *R*.

In order to avoid any possible misinterpretation, we emphasize again that it would be inappropriate to cite this study as a justification for using two-item scales. True-score theory indicates that, all other things being equal, more items lead to better construct representation and the primary way to make measures more reliable is to increase the number of items (Herbert et al. 1998; Emons et al. 2007). If, however, research design or off-design circumstances dictate that the scale has two possibly congeneric items, then it is best to report the Spearman-Brown reliability estimate.

Acknowledgments The authors are grateful to William Revelle and an anonymous reviewer for helpful comments on a previous version of this manuscript and suggestions for improvements.

References

Bollen KA (1989) Structural equations with latent variables. Wiley, New York

Cramer ME, Atwood JR, Stoner JA (2006) Measuring community coalition effectiveness using the ICE© instrument. Public Health Nurs 23:74–87

Cuijpers P, Smits N, Donker T, ten Have M, de Graaf R (2009) Screening for mood and anxiety disorders with the five-item, the three-item, and the two-item mental health inventory. Psychiat Res 168:250–255

Embretson SE, Reise SP (2000) Item response theory for psychologists. Lawrence Erlbaum Associates, Mahwah

Emons WHM, Sijtsma K, Meijer RR (2007) On the consistency of individual classification using short scales. Psychol Methods 12:105–120

Hancock GR, Mueller RO (2001) Rethinking construct reliability within latent variable systems. In Cudeck R, Du Toit S, Sörbom D (Eds), Structural equation modeling: present and future. A festschrift in honor of Karl Jöreskog. Scientific Software International, Lincolnwood, pp 195–216

Herbert W, Marsh HW, Hau K-T, Balla JR, Grayson D (1998) Is more ever too much? The number of indicators per factor in confirmatory factor analysis. Multivar Behav Res 33:181–220

Hulin C, Netemeyer R, Cudeck R (2001) Can a reliability coefficient be too high? J Consum Psychol 10:55–58

Little TD, Lindenberger U, Nesselroade JR (1999) On selecting indicators for multivariate measurement and modeling with latent variables: when 'good' indicators are bad and 'bad' indicators are good. Psychol Methods 4:192–211

Lord FM, Novick MR (1968) Statistical theories of mental test scores. Reading, Addison-Wesley

Löwe B, Kroenke K, Gräfe K (2005) Detecting and monitoring depression with a two-item questionnaire (PHQ-2). J Psychosom Res 58:163–171

Michal M, Zwerenz R, Tschan R, Edinger J, Lichy M, Knebel A, Tuin I, Beutel M (2010) Screening for depersonalization-derealization with two items of the Cambridge depersonalization scale. Psychother Psych Med 60:175–179

O'Brien M, Buikstra E, Hegney D (2008) The influence of psychological factors on breastfeeding duration. J Adv Nurs 63: 397–408

Revelle W, Zinbarg RE (2009) Coefficients alpha, beta, omega, and the GLB: comments on Sijtsma. Psychometrika 74:145–154

Rüsch N, Corrigan PW, Wassel A, Michaels P, Olschewski M, Wilkniss S, Batia K (2009) A stress-coping model of mental illness stigma. I. Predictors of cognitive stress appraisal. Schizophr Res 110:59–64

Sainfort F, Booske BC (2000) Measuring post-decision satisfaction. Med Decis Mak 20:51–61

Sijtsma K (2009) On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 74:107–120

Verhoef PC (2003) Understanding the effect of customer relationship management efforts on customer retention and customer share development. J Mark 67:30–45

Young J, Jeganathan S, Houtzager L, Di Guilmi A, Purnomo J (2009) A valid two-item food security questionnaire for screening HIV-1 infected patients in a clinical setting. Public Health Nutr 12:2129–2132