Introducing Consistency Scales in Survey Instruments

Gerard J. Fogarty

Centre for Sustainable Business Development, University of Southern Queensland

Toowoomba, QLD 4350

Fogarty@usq.edu.au

&

Nicole Steele

Directorate of Mental Health, Department of Defence

Canberra, ACT

The correct APA citation for this paper is:

Fogarty, G. J., & Steele, N. (2013). Introducing consistency scales in survey instruments. In R. Lacey (Ed.), *Advances in Organisational Research Methods and Analysis*, pp. 144-155. New York: Chamber & Row.

Abstract

Major psychological test instruments, especially the longer ones, often contain embedded validity scales. The intent of validity scales is to detect individuals who may be presenting a distorted picture of themselves either by deliberately faking responses or by responding to the items without understanding their meaning or perhaps by simply not reading the items and responding randomly. Different types of validity scale are constructed to target each of these response patterns. The response pattern of concern in this chapter is random responding and the relevant validity checks are usually referred to as consistency scales. For example, the item “I find my job stressful” should elicit a similar response to the item “There is a lot of stress in my job”. A pair of dissimilar items, on the other hand, should elicit responses that are in the opposite direction. Organizational psychologists know about these scales but they tend to neglect them when constructing surveys. This chapter presents a case study that illustrates the methodology involved and the effect of developing and implementing consistency checks in surveys.

Background

Despite its pitfalls, psychological assessment in organizational settings continues to rely heavily on self-report methodology. It is efficient, convenient, and often the only means of gathering information about psychological constructs of interest to employers, trainers, managers, and staff. However, this form of assessment is plagued by two major problems: impression management and response inconsistency.

Regarding the first of these, response distortion in the shape of *faking good* and *self-deception* is a major threat to the validity of self-reported assessments, especially in the personnel selection field where individuals are likely to be motivated to convey a favourable impression. Much of the interest in response distortion has been driven by widespread use of personality tests in selection settings. The journal, *Personnel Psychology*, recognised this level of interest when it published a series of articles by prominent organisational psychologists outlining the pros and cons of the use of personality tests for personnel selection (Morgeson et al, 2007a, 2007b; Ones, Dilchert, Viswesvaran, & Judge, 2007; Tett & Christiansen, 2007). Even allowing that there were different points of view expressed in this debate, there is no denying the seriousness of an issue that leads some experts to claim that the response distortion problem is intractable and that self-report measurement of personality should therefore be abandoned (Morgeson et al, 2007a).

Because it is not our intention to write at length about response distortion, we will not cover this debate but we do note the prominence of the issue. We also note that concern about distortion is just a part of a more wide-ranging concern about response styles that threaten the validity of self-report instruments. Major psychological test instruments, especially the longer ones, often contain embedded validity scales. The MMPI-2 (Butcher et al, 2001), the PAI (Moray, 1991), and the NEO-FFI (Scandell, 2000), are examples. The intent of validity scales is to detect individuals who may be presenting a distorted picture of themselves either by deliberately faking responses *or by responding to the items without understanding their meaning or perhaps by simply not reading the items and responding randomly*. Different types of validity scale are constructed to target each of these response patterns. The response pattern of concern in this chapter is random responding and the relevant validity checks are usually referred to as consistency scales.

Response Inconsistency

Interestingly, whereas response distortion occurs in situations where individuals have a stake in creating a good impression, response inconsistency tends to occur in situations where the motivation to respond is not high. Organizations that regularly use surveys to assist in organizational improvement initiatives are likely to find that a proportion of their employees either do not respond at all or respond in a haphazard, half-hearted way. Researchers and managers are aware of the non-responders because data are missing. Statistical textbooks that deal with data screening give cautionary advice about missing data, suggesting ways of replacing missing data according to whether it is missing on a random or a non-random basis. Non-random missing data are usually associated with variables where there is some reason why people have not responded. Random missing data are harder to explain but at least you have the advantage of seeing that the data are missing.

Far harder to detect are responses that have been made without due thought and consideration. Anyone who has ever hand-scored an organizational survey with a large number of items knows that not all respondents read every question carefully before marking their responses. Consistency scales can help to detect these people, yet they are rarely included in organizational surveys. In this chapter, we illustrate how a consistency scale can be developed and implemented in an organizational climate survey.

Organizational Context

The *Profile of Unit Leadership Satisfaction and Effectiveness (ADF PULSE*: Goyne, Riley, & Johnston, 2008) grew out of a need among the Canadian Forces (CF) and the Australian Defence Force (ADF) to assess organizational climate in a garrison environment. Scales measuring workplace demands, motivation, satisfaction, performance, teamwork, communication, commitment, support, and job intentions were developed, mostly via the adaptation of existing instruments. A comprehensive demographics section includes items measuring exercise routines, drinking and smoking habits, and deployment history. With over 200 items and associated measures spread over 12 pages, the ADF PULSE is a reasonably large survey, certainly long enough to warrant the inclusion of a consistency scale. The Commander of a relatively small unit does not want data corrupted by a small number of fatigued or uninterested individuals.

The Rationale Underlying Consistency Scales

There are different ways of constructing consistency scales. We will deal with one of the easiest and most common methods. The Variable Response Inconsistency (VRIN) scale from the MMPI was the model used for the construction of the ADF PULSE consistency scale. The VRIN consists of item pairs that have similar or opposite meanings. A pair of similar items should elicit similar responses. For example, the item “I find my job stressful” should elicit a similar response to the item “There is a lot of stress in my job”. A pair of dissimilar items, on the other hand, should elicit responses that are in the opposite direction. For example, the item pairing: “I wake up fresh and rested most mornings” and “My sleep is fitful and disturbed”.

The logic underlying the use of consistency scales is that if an individual responds in an inconsistent fashion enough times, there is good reason to suspect the validity of that person’s data. In a clinical setting, this information could affect the interpretation of the results for the self-report instruments used. In an organizational setting, the case might be deleted from the dataset before proceeding to analyse means and relationships among variables.

Constructing the ADF PULSE Consistency Scale

There are two ways of identifying item pairs to include in a consistency (or inconsistency) scale. The first way is to deliberately embed items that will attract similar or opposite endorsement patterns. LePage, Mogge, and Garcia-Rea (2009) took this approach with the short Assessment of Depression Inventory (ADI). One of the drawbacks of this approach is that it may involve the inclusion of items that have little to do with the constructs measured by the scale. A second approach is to analyse item inter-correlations and to select pairs with high positive or high negative correlations. This is the most common methodology and one that is well-suited to longer instruments, such as the ADF PULSE. That was the approach taken here. An ADF PULSE master database (N = 3,596) was used to calculate inter-item correlations. Pairs of items were selected on the basis of seven criteria:

1. The first criterion was that the items be substantially correlated, either positively or negatively. Instruments like the MMPI use item pairs with correlations above .70. LePage et al. (2009) showed that when inter-item correlations are as high as ± .70, a set of four items is sufficient for a consistency scale. However, the ADF PULSE does not contain that degree of item redundancy. In particular, whilst there are numerous instances of high positive

correlations among items, there are not many instances of high negative correlations, so some of the reverse-direction item pairs were based on correlations as low as -.30. Lower inter-item correlations mean that more items are needed to form the consistency scale. To compensate for the lower inter-item correlations, the ADF PULSE consistency scale contains 25 item pairs.

2. The second criterion was that the members of the pair look as though they should elicit same- or opposite-direction responses. This criterion was applied because it was not always possible to see why two items would have a substantial positive or negative correlation. When this situation occurs, there is always a suspicion that the true correlation may be less than the observed and that the high correlation in the base sample may not prove reliable over time. Thus, if two item pairs had similar correlations, this second criterion was used to choose the pair with the plausible correlation in the belief that the relationship would prove more reliable across a range of samples and contexts.

3. The third criterion was that the members of the pair are not too near each other in the survey. This criterion was applied because individuals responding in a random fashion are likely to notice similarities or dissimilarities between adjacent items. Pairs that contain widely-separated items are more likely to be sensitive to random responding. We note, however, that it is not always possible to apply this principle.

4. Response inconsistency is a complicated topic and it is likely that different causes underlie inconsistent responses to same-direction items compared with inconsistent responses to opposite-direction items. For example, someone who agrees with most items or disagrees with most items will inevitably end up with a high consistency score if the scale contains only same-direction item pairs. However, that person would obtain a very low score if the scale contains only opposite-direction item pairs. Accordingly, the fourth criterion was that every attempt was made to select an equal number of same- versus opposite-direction item pairings.

5. The fifth criterion was that item pairs be sampled from the beginning, middle, and end sections of the ADF PULSE survey to check for signs of survey fatigue. If there are enough items in the consistency scale, as was the case here, you can end up with Consistency sub-scales. Thus, a respondent who started out enthusiastically but then lost interest in the latter stages of the survey may end up with a reasonable consistency score overall but a low score for the sub-scale corresponding to the final section. Should that be the case, a reasonable proportion of that respondent’s data are probably usable.

6. The sixth criterion was that no item should appear in more than one pair. This is not what we would call a hard-and-fast criterion and it may be that there are so few high correlations among pairs of items that you are forced to use a “good” item more than once. In fact, this is the situation we faced with ADF PULSE.

7. A seventh criterion was that items were taken from sections of the survey that were relevant to all respondents. If this principle is not applied, some adjustment will be necessary for respondents who cannot complete some sections of the survey.

Sample item pairs from the ADF PULSE consistency scale are shown in Table 1.

Table 1. Consistency Items from Different Sections of ADF PULSE

Item	**‘r’**	**Section**	**Description**
1	**-.32**	Opening	The priorities of my work are clear to me I have conflicting priorities at work
4	.66		I find my work inherently rewarding My work fits my interests and skills
6	**-.59**		I am not satisfied with the pay and benefits I receive I feel I am being paid a fair amount for the work I do
8	.74		I like doing the things I do at work I am satisfied with the kind of work I do in my current job
10	**-.49**	Middle	Commanders set the example for compliance with standards Unit leaders allow the cutting of corners to get a job done
14	**-.58**		I enjoy being part of the social activities of my work group My workgroup members rarely socialise together
16	.50		I like the people I work with I would miss members of my work group if I was to stop working with them
17	.61		My work group is united in trying to reach its goals for performance I like the work practices of my work group
20	**-.51**	End	My unit does not appreciate any extra effort from me The unit takes time to recognise my achievements
22	.54		My unit cares about me The unit treats me as a responsible person

Note: There were 25 item pairings in total

Scoring Rules

There are various techniques for scoring consistency, even within the methodology we have chosen here. The rules we used to score the item pairings were as follows:

- The response format for ADF PULSE items employed a Likert format: Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree. If the two items were in opposite directions, a score of 1 was registered if individuals agreed (Agree, Strongly Agree) with one item and disagreed (Disagree, Strongly Disagree) with the other. SPSS syntax was used for these calculations.

- If the two items were in the same direction, a score of 1 was registered if individuals agreed (Agree, Strongly Agree) with both items or disagreed (Disagree, Strongly Disagree) with both. Otherwise the score for that Consistency item was zero.

- The resulting 1's and 0's were then added to yield a score out of 25. Scores were converted to percentages. A score of 100% suggested that individuals read all questions carefully and responded thoughtfully.

Determining a Cut-Off Score

This method of scoring makes it easy to calculate the probability of obtaining a correct response on a purely random basis. This can be done by looking at the various response combinations and calculating the proportion that would yield a score of 1. The proportion is 68%. Simulation methods can also be used to obtain a theoretical distribution. Using simulation methods, a set of 300 randomly-generated responses (around the maximum sample size for an ADF PULSE survey) yielded a mean Consistency score of 67.7% and a standard deviation of 9.4%. The standard deviation can be used to set a cut-off value that would exclude a certain proportion of the population (e.g., 1.64 SDs below the mean). However, this is not the method we recommend.

A different cut-off point is obtained if one analyses the actual Consistency scores obtained by ADF PULSE respondents in the base dataset (N = 3,596). The average consistency score was 91% (SD = 8.75), which is a very high figure indeed, suggesting that the respondents were motivated to complete the ADF PULSE instrument. The random distribution and the actual distribution of Consistency scores are shown in Figure 1.

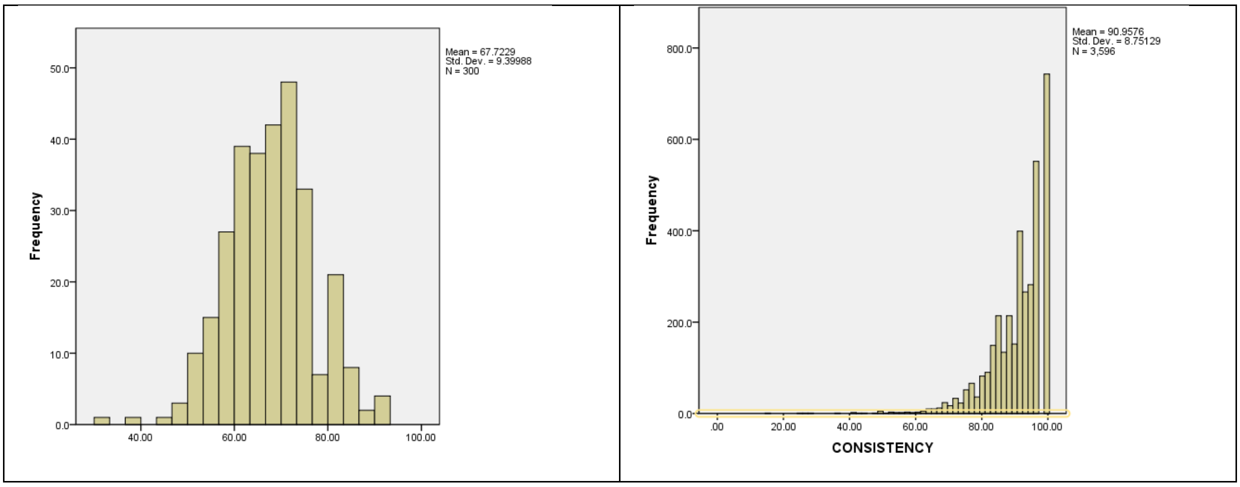

Figure 1. Distribution of Consistency scores for random (left) and actual (right) responding

Using the actual distribution (rhs of Figure 1) as the basis, a cut-off score of 76% (1.64 SDs below the mean) would exclude the bottom 5% of respondents.

A third possible cut-off value was suggested by practices adopted by publishers of large instruments, such as the MMPI, where it is not uncommon to choose a point two or even three SDs below the mean. A point two SDs below the mean for the ADF PULSE would result in a cut-off score of 73.5%.

Any scores below the cut-off mark should trigger an inspection of that individual’s data before a decision is made about excluding the case. We will say more about this matter when we look at the sub-scale scores.

Analysis of Consistency Scores

As mentioned above, the consistency items in ADF PULSE were selected to cover the beginning, middle, and end sections. Examining the scores across the three sections helps to decide whether respondents were being inconsistent throughout the survey or in particular sections. There was a significant decline in consistency scores across the beginning, middle, and end sections of ADF PULSE, suggesting that fatigue may have become a factor towards the end of the survey. The trend is illustrated in Figure 2.

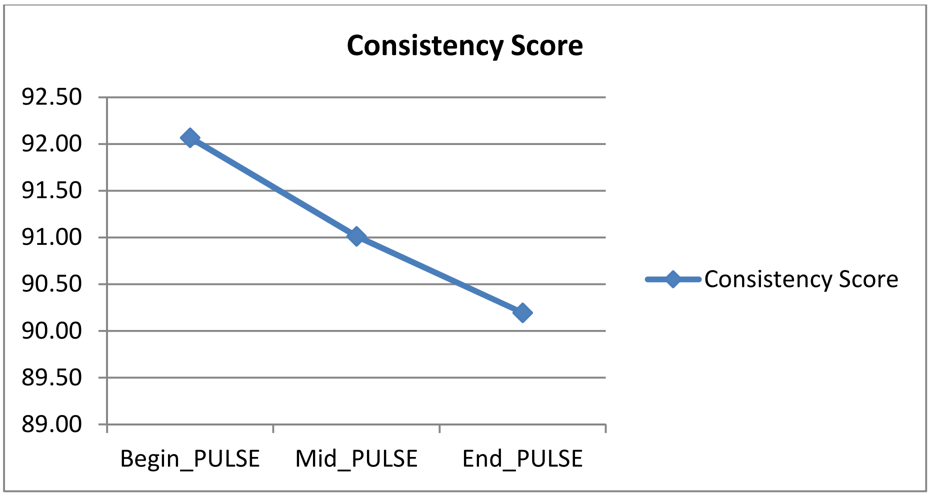

Figure 2. Decline in consistency scores across ADF PULSE sections.

Although statistically significant, the decline is not dramatic, reinforcing the impression created by the high mean score that these respondents were committed to the task. On an individual basis, however, it is likely that some respondents have begun to respond in at least a partially random fashion by the end of the survey, emphasising the need to inspect low scorers on a case-by-case basis.

To assist in this case-by-case inspection, SPSS syntax can be used to produce a list of individuals who fail to reach the cut-off score. Sample output is shown in Table 3.

Table 3. Example of SPSS Output for a 2010 ADF PULSE Sample

ID	Total C	BeginningC	MiddleC	EndC
26	25.32	47.50	10.00	25.00
95	64.94	90.00	80.00	25.00
113	71.43	60.00	60.00	90.00
141	48.70	67.50	43.33	37.50

This extract from the SPSS output shows how easy it is to check individual cases. The ID is shown in the first column, the total Consistency score expressed as a percentage in the second column, and the three section scores in the remaining columns. It is not unusual to see wide variation in scores across the sections. The task then becomes checking each of these cases to see whether they should be deleted or whether part of their data can be used. The first respondent (ID 26) had rather poor consistency scores throughout and is therefore a candidate for deletion. The second respondent (ID 95), who would probably have been deleted on the basis of total score, had good consistency scores for most of the survey but dropped off badly in the last section.

Relationship with Other Variables in ADF PULSE

The consistency scale is a validity scale but because it assesses motivation and thoroughness, we considered it worthwhile to examine its relations with other variables. There was a significant difference between ranks with higher ranks showing greater consistency. Consistency was negatively correlated with most of the "undesirable" variables in ADF PULSE (e.g., stressors, burnout, K10) and positively correlated with all the "desirable" variables (e.g., job satisfaction, communication, safety, job performance). In other words, people who responded consistently tended to have better psychological profiles, suggesting that the decision to respond randomly or to not answer questions was partly driven by psychological reasons as well as a desire to complete an assigned task in the shortest possible time.

Caveats in the Construction of Consistency Scales

Perhaps we have made it sound overly easy to construct consistency scales, so we close with a few caveats. Firstly, achieving a balance of same-direction and opposite-direction pairings is important but will not be possible if all items are positively-oriented. Paradoxically, one of the main reasons for using reverse-coded items in surveys is to encourage respondents to read the items closely, and thereby improve consistency. However, a mixture of reverse-coded items and normal items often leads to situations where the two types of items end up defining separate factors (Marsh, 1996). To avoid this unwanted problem, survey developers sometimes avoid reverse-coded items altogether or these types of items are removed in revisions of the survey instrument that aim to improve internal consistency reliability estimates

The end result is that it is not uncommon to find organizational surveys made up entirely of positively-worded items, in which case it will not be possible to compose opposite-direction pairings for the consistency scale. Consistency scales can still be constructed in this situation but the method described in this chapter would not detect people who always selected response options from the same end of the response scale (e.g., agreed with all items or disagreed with all items).

Secondly, it is our experience that consistency scales are of more value when the results of the survey may not be of concern to all the respondents; a situation which covers a great deal of the climate survey work currently conducted in organizations. To illustrate this point, in another context the first author constructed a consistency scale for a safety climate instrument that formed part of a job selection test battery. The mean score on the Consistency scale of respondents in that situation was 88.6%. The same safety climate survey was administered to university students as part of a project on road safety. Students may have had an interest in the topic but it is reasonable to assume that most of them would have completed the 80-item survey for course credit. In this situation, the mean score on the Consistency scale was 66.1% with many respondents failing to reach the cut-off point.

Outcomes of this nature suggest that consistency scales may not be worthwhile in selection situations and that they are more useful in situations where the stakes are low for the individual. But we would not wish to generalise to that extent. In a selection situation, a validity scale that identifies even 5% of the test forms as requiring further investigation may still be a valuable aid to selecting the right candidates for the job.

Recommendations

We recommend the introduction of consistency scales wherever possible because:

a. Consistency scales are data screening devices and deleting cases where there is strong evidence of inconsistent responding improves the quality of the data to be interpreted;

b. They enhance the professionalism of the service the survey administrator is offering;

c. They are efficient in the sense that once the scales have been constructed, scores can be computed using exactly the same statistical packages or spreadsheets that are used for the other scales in the instrument.

As we have shown here, Consistency is an interesting variable in its own right, demonstrating positive correlations with positive traits (e.g., job satisfaction) and negative correlations with negative traits (e.g., Stress).

References

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *MMPI-2 (Minnesota Multiphasic PersonalityInventory—2): Manual for administration, scoring, and interpretation,revised edition.* Minneapolis: University of Minnesota Press.

Goyne, A., Lake, R., Riley, M., Johnston, B. (2009). Taking the PULSE of your unit: A command support tool for assessing unit climate. In P. Murphy (Ed.), *Focus on human performance in land operations*, (pp. 70-77). Department of Defence, Canberra, Australia.

LePage, J. P., Mogge, N. L., & Garcia-Rea, E. A. (2009). Detecting random responding using the Assessment of Depression Inventory: A brief screening measure of depression. *Depression and Anxiety, 26*, 592-595.

Marsh, H. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*(4), 810-819.

Morey, L. C. (1991). *Personality Assessment Inventory*. Odessa, FL: Psychological Assessment Resources.

Morgeson, F.P., Campion, M.A, Dipboye, R.L., Hollenbeck, J.R., Murphy, K., Schmitt, N. (2007a). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*, 683–729.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60*(4), 1029-1049.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*(4), 995-1027.

Scandell, D. J. (2000). Development and initial validation of validity scales for the NEO-Five Factor Inventory. *Personality and Individual Differences, 29*, 1153-1162.

Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007). *Personnel Psychology*, *60*(4), 967-993.