

## Response Patterns and Impression Management

International Test Commission Conference –

Winchester, UK June 2002

***Norman Buckley & Rebekah Williams, Redfield Consulting***

This conference is seminal in the field of internet based testing. We are hearing of powerful technical developments and there are very germane points being made of a theoretical, logistical and ethical nature. However I would like to look at some of the simple, practical issues associated with moving from paper and pencil to web-based data capture. To do this I will describe what we did when we moved a proven, reliable, paper and pencil Big 5 questionnaire to the web. I will discuss some of the issues and questions that came up and some of the techniques we created and implemented to address them.

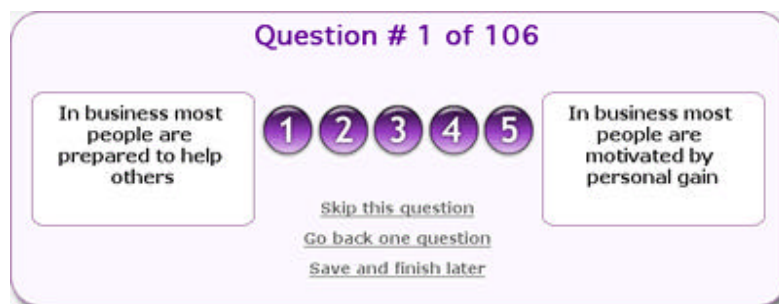
In particular I will describe an approach to monitoring Impression Management that we believe holds great promise.

Facet5 is a Big5 questionnaire. It has been around since the late 1980's and although computer admin was always an option, most people continued to use paper data capture and then computer scoring and reporting. We had a lot of pressure from clients to change our approach. A real issue was that of "foreign" software. Companies are getting more and more cautious about having third party software on their systems. We were told quite forcefully that organisations wanted a completely web-based system with no need to download anything other than, perhaps, completed reports in pdf format. So we set about the conversion. Many technical problems were met – possibly one of the most intractable being the process for producing the formatted reports. However these were all resolved.

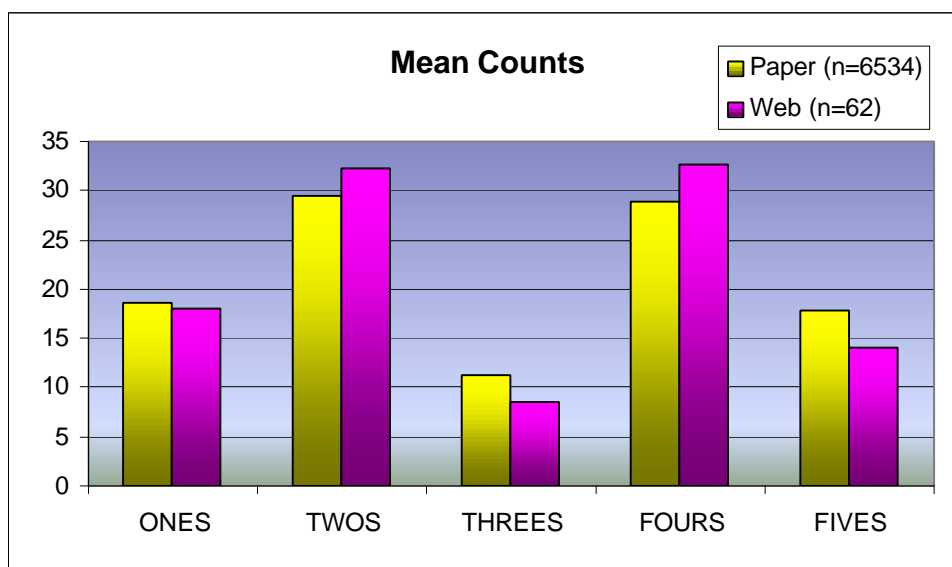
However a core question that emerged was equivalence. Moving to the web changes the format for both administration and presentation and we are admonished to be very careful when we change such aspects since the effect on the scores is unknowable. Therefore when we moved Facet5 to web based administration it was

important that we determined whether the responses we got were comparable with those we had been getting from paper and pencil.

What do we mean by comparable? While this can be a complete study in its own right, I want to focus here on one aspect: Response Patterns. Facet5 consists of 106 semantic differential items. The semantic differential format was chosen since it is believed to give quicker response times than the traditional Likert scale, possibly since both the statement and its implicit antonym are given. There is less need for a person to interpret the antonym for him/her self. It looks like this.



We would expect this format to produce a bimodal distribution with most answers falling onto 2 and 4. If web-based administration was going to have a significant impact on the response pattern it seems likely it would show up first here with a pattern that was significantly different. Figure 1 shows the Response Distributions for both paper and web based questionnaires. It can be seen that the patterns are broadly similar - none of the differences are statistically significant.



From this we can assume that Facet5 does not produce substantially different results when presented on the web than it does from paper & pencil - apart from proving quicker. On paper Facet5 was generally found to require 20-25 minutes to complete. Web presentation has reduced this to 16-17 minutes.

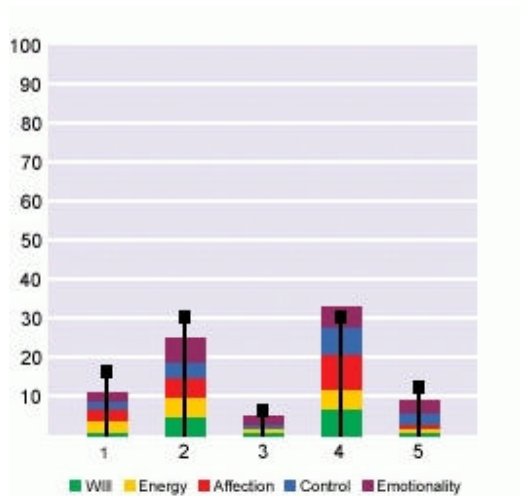
However the second part of the question is much more significant – can we trust the data?

Impression Management has been a major question ever since non-cognitive tests were created. It is felt that, under many circumstances, people will be motivated to answer in a way that produces a profile that fits a “perceived” purpose rather than being a true reflection of themselves. This has implications for interpretation, especially where important decisions might be made on the basis of the results. Selection for jobs is an obvious such application and probably the most widely used. However there are others such as:

- Counselling – people may exaggerate or suppress aspects of a profile
- Clinical diagnosis – people may attempt to present symptom sets for a specific purpose.
- Forensic – people can have a very strong motive for presenting one set of elements over another.

So what can we do? First we need to define what we mean. We must differentiate between distortions produced by “natural” biases (extreme responses, central tendency, random responding, etc), which would be expected to affect all scales equally, exaggerating or flattening a profile, and deliberate IM where a person is trying to produce a profile to fit a particular purpose. I would like to call them Response Bias and Impression Management respectively. Ways of identifying Response Bias tend to be somewhat mechanical. Response patterns are compared to reference sets and little advice is available to aid interpretation. In Facet5 our biggest concern has been central tendency (an over-use of the middle response). Extreme response patterns are really quite rare. Our test admin module now monitors the distribution in real time and compares the final distribution to a reference one. If there is a severe distortion (far too many “3”s for example) the “offending” items are represented at the end and the respondent is given the chance

to re-think. The final decision is obviously accepted. We then report this in graphic



form as follows:

Here the black bars show the expected frequencies (based on a sample of approximately 6000 cases) and the respondents answers are shown with the different factors differentiated by colour. This can be very helpful in discussion if the "3"s are predominantly in one domain. We are now implementing the ability to click on the "3" bar and see which items are causing the indecision.

IM in the other hand is much more widely researched and can be quite subtly managed. Impression Management (IM) is an attempt (not necessarily deliberate) to produce a profile that is different from the respondent's "true" or "natural" profile and in-line with a perceived expectation. There are three broad types of IM that might occur:

1. Denial, Defensiveness or Suppression - attempting to suppress anything that might be perceived as negative. This may be identified by specific "marker" questions or by the proportion of "No" (denied) responses or by some combination of the two. It is likely that a candidate for a job would try to present the best possible self-portrait so this type of IM is more prevalent in selection than development.

Indicators of this type of IM may be called Social Desirability, Motivational Distortion or Defensiveness. High scores on these scales are supposed to act as

warning bells suggesting caution in interpreting the results. The person may not be being "open" or "honest".

2. Suggestibility, Faking Bad - this is the logical opposite of the first. A respondent may exaggerate faults or over-admit to possible problems. Such reactions are not uncommon in clinical cases (a cry for help?) but can also be seen in areas such as Career and Forensic psychology.

Such a response pattern is rarely picked up by marker questions. Very low Social Desirability or Motivational Distortion scales are usually seen as being "open" or "honest". Scales using a "Yes" or "No" format can identify such Suggestibility by a disproportionate number of "Yes" responses. Again caution is urged in interpretation.

3. Templating - where a respondent has a mental image of the "ideal" profile and tries to adjust responses to match this supposed "ideal". This is the most likely type of IM in selection and yet typical Social Desirability or Motivational distortion Scales do little to identify it. Psychometric folklore is littered with statements from people who claim they can manipulate questionnaires to present any picture they want. My bank manager announced, after I told him what we did for a living, *"Oh yes I know those things, I always come out really well but you can make them say anything you like can't you?"*

Test developers have adopted a number of strategies to minimise the likelihood of IM and to identify it if it does occur. Some techniques include:

- Item ambiguity - word the items so it is not obvious which factor they load on. This is not always easy to achieve and failure to do so can affect the results obtained. For example the 16PF uses ambiguity to a certain degree but some of the items loading on Emotional Stability and Self Assurance (Factors C and O) are fairly transparent. (Q94 – *"I am troubled by guilt or remorse over quite small matters"* or Q119 – *"I occasionally have periods of feeling depressed, miserable and in low spirits for no apparent reason."*). In a selection application few people will mark such items so that average scores are considerably below that found in the general population. These two scales are the heaviest loadings for the second order factor Anxiety and Bartram<sup>1</sup> showed that the mean sten score for Anxiety was about 3 in a sample of short-listed candidates.

- Neutral or balanced valence - make it hard to identify which response, representing opposite aspects of a scale, is more desirable. This is not the same as ipsativity where the respondent is required to choose between statements representing different domains (some OPQ versions, DISC etc). Balanced valence uses pairs of statements that represent opposite ends of a single scale. They may be arranged as separate statements (MBTI) or semantic differentials (Facet5).
- Marker items - a group of items is included, which, it is assumed, represent the behaviour of "normal" people. These items will include some that admit negative characteristics on the assumption that most people have some "bad" elements in their make-up. Over denial of these items is seen as an attempt to appear unrealistically "good".
- Selective norming - responses are compared to other people in a similar position eg. applying for a job in retail sales. Here all applicants are expected to have the same degree of motivation to distort and so the effect of IM will be cancelled out. However this assumes that all respondents will apply the same IM strategies to the same degree. It also raises the spectre of having to re-norm the profile when the person is hired. How do you then explain that the score on a scale was 7 when an applicant but it is now 9 as an employee?

Most developers use the first two methods to reduce the effect of IM. Some include the third but the fourth is less common. The collection of multiple norm tables is technically simple but logically questionable and practically problematic.

The net result of all this is that, if a high MD, SD or Faking scale accompanies the profile, we are advised to exercise caution in interpretation. Some questionnaires suggest scale adjustments, especially those that have a more clinical origin (the criterion referenced ones such as the MMPI, CPI and Humm Wadsworth stand out). The 16PF has compensation algorithms but Cattell<sup>ii</sup> in his discussion of the impact of Motivational Distortion urges great caution in interpreting MD measures describing them as a "*Temporary Compromise*" while "*more basic research proceeds.*" (p55). He warns that automatic adjustment using such measures will by definition "*take out real personality variance as well as motivational shift*". (p56)

It is known for example that MD and SD scores tend to correlate with elements of conscientiousness and empathy. In the 16PF people with high MD scores are seen as Warm-hearted, Happy-go-lucky, Venturesome, Emotionally Stable, Unperturbed, Relaxed, Conscientious, Practical, Self-sufficient, Controlled, and Trusting. In terms of second order factors broadly aligned with the Big5, these are Stable, Conscientious Extraverts. It is also highly likely that they will produce high MD scores.

So these IM scales might tell us that a profile is suspect but:

1. It may not be and
2. We get little advice as to what to do with the information.

So what do we do? Lets go back to first principles. Put yourself in the position of the respondent for whom it is important that the right picture is presented. I suspect this applies in most applications but I'm going to focus on Selection, which is an area I know. Lets take, as an example, a music TV Company. This company operates in the trendy, youth market. It presents a radical, cutting edge image in everything from it's TV channels to its website, corporate literature and Central London premises like something out of Red Dwarf. Right from the start the company is projecting its culture clearly and unambiguously. A candidate will form an image of what it is like and, from this infer what is expected.

As part of the recruitment process they are then presented with a sequence of questions, some of which appear, to them, to be tapping in to the core values of the company. So what happens? Let's say we have items tapping a "Conservative-Radical" dimension (not hard in a Big 5 structure). Everything so far suggests that the company is towards the "Radical" end of this scale. The candidate doesn't want to appear too "Conservative" so will tend to think before answering these type of questions. They may moderate their responses slightly to take account of their perception of what the company wants. This is IM at work. But what does this mean to their scores? Do they "nudge" them up or down? Frankly we don't know. It probably depends on many factors including:

- How important is the job to them?
- How malleable is the respondent?

- How clearly did the organisation telegraph its values and style?

So we don't know. But what we do know is that they thought about it. They probably thought more about it than about the other questions that they did not perceive were related to the "important" domains. So this gives us a clue where to look.

I want to divert a little here to look at a broader perspective. It's probably no surprise but it appears that while we were following what we thought was an original and promising theoretical path, others were coming to similar conclusions. Forensic psychologists have obviously been concerned for some time about how to detect lying and other economies of truth. Little snippets have been emerging from time to time and Aldert Vrij summarised a number of them in the *Psychologist* in November 2001. These include:

- Most stereotypes of "lying" behaviours are false
- Even trained police aren't very good at identifying lying
- Really good liars are very hard to spot
- Liars take a little longer before responding!!!

This may be linked to increased "cognitive load". i.e. a person is doing more than just answering the question; they are trying to judge the "best" answer to the question.

So the threads converge. I hasten to add that it might be a bit much to equate a person who really wants a particular job and wants to put their "best foot forward" with the purloiner of a truckload of stereos. However some of the same forces are at work. The person is thinking about what to say – they are under higher "cognitive load".

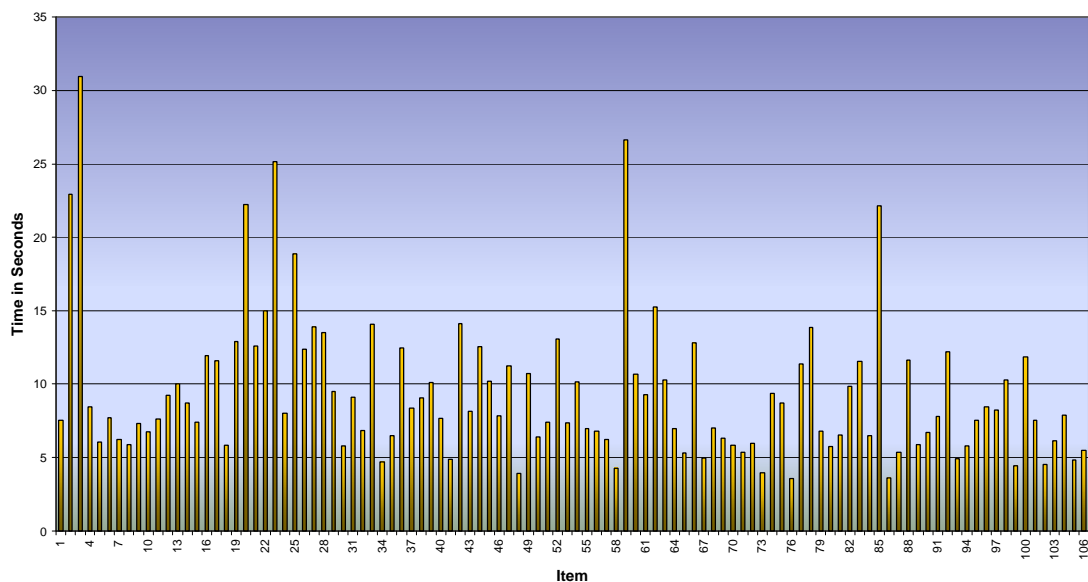
So what can we do about it? How can we use this insight? Forensic psychologists might, for example, structure a series of questions covering multiple domains including but not limited to those thought to relate to the crime. With the benefit of post-interview analysis the time taken to respond to questions can be determined and we can see whether "Light Fingered Louis" takes longer on average when asked

about “Friday the 14<sup>th</sup>”, the HiFi Warehouse, big yellow trucks etc than when asked about the menu at his sister’s BBQ.

IO psychs can do the same. Structure a set of questions, some of which might be perceived to be key to the role. Measure the time taken to respond and check to see whether there is a pattern in the items which Louis took extra time over. This is exactly what we did with Facet5.

When we created the database for the web version of Facet5 we added the ability to record Response Latency – how long did the person take to respond. Typically we found that people take about 8 seconds to respond to a single question. But there is a big range. Here is a typical set of response latencies.

**Facet5 Item Response Times**



A spike on Item 1 is very common. We suspect it is due to people taking a few seconds to settle. There are also other spikes, perhaps due to a call of nature, urgent need for coffee, emails arriving etc. However they are obvious and can be filtered out.

We can then sort the remaining items from quick to slow and split off the slowest 15 or so. Now this is where the theory gets interesting. If a person has little interest in presenting a favourable or “managed” result, then no items will stand out over others and the slow ones will be evenly distributed across all 5 domains. However if

someone has focused on one domain over another (eg thought a lot about the Radical-Conservative one) then items from that domain should be over-represented in the "Slow" set.

We did not have access to large research samples so to test this theory we had to "hunt in the wild". Facet5 has been in use on the web since January 2001 in a number of settings. We were demonstrating it extensively, mostly to groups of psychologists and other professionals in the HR field. Here the objective was to show what Facet5 looked like and its capabilities. Many people asked to trial it themselves. They were all people who were very familiar with other processes of varying complexity (MBTI, CPI, OPQ etc). Their objective was to see what Facet5 looked and felt like and how it compared to the other processes they were used to. One might assume that the need for IM was low in this group.

However Facet5 was also being used in earnest. A number of users were applying Facet5 as part of an assessment process used to select people for real jobs. It is reasonable to assume that these people would want to present themselves in as good a light as possible so the drive to IM was higher.

So this gives us two groups thought to differ in levels of IM. So how do we test the difference? This seemed simple. The group of professionals and psychologists (lets call them Group A) would be expected to show "slow item distributions" that were even across all five domains. The real applicants (lets call them Group B) should have slow item distributions that are not even. A simple way of assessing this would be to calculate  $D^2$  distributions for the two groups and see if they differ.  $D^2$  was calculated in the traditional way as follows:

$$D^2 = \sum_{i=1-n} (X_i - 3)^2$$

where X are the factor scores for person X and 3 is the expected value for a "perfect distribution" i.e. IM=0. The results were as follows:

	Group A (low IM)	Group B (high IM)
N	<b>17</b>	<b>18</b>
<i>MeanD<sup>2</sup></i>	<b>7.05</b>	<b>10.47</b>
<i>SD</i>	<b>4.19</b>	<b>3.57</b>
<b>T = 0.01 df=35</b>		

So bingo – it seems to work. When you have a strong desire to present in a particular way or “looking good”, your RLA’s become uneven. We believe that this is because a greater amount of “cognitive load” is being applied to items “perceived” to be salient.

So how does this help us?

- Like traditional IM scales we have information that suggests that IM may be present.
- Like traditional IM scales we can urge caution in interpretation in certain cases.
- We can’t tell you how to adjust the scores you have got to take this IM into account but this is probably not a good idea anyway.

But what we can do is to tell you exactly where the IM is focused. This is an immensely powerful aid to interpretation. It is the first point raised during feedback. Then people will tell you:

- Yes I’m probably not that disciplined at home
- I was trying to differentiate between me at work and me at home
- There were some items where I truly was in the middle.

A revision is under way now to allow users to not only see that a number of items caused Louis to think hard but one click will show what they were. Suddenly we are encouraging respondents to explain and explore the thinking behind their response patterns.

Where to from here?

1. Replicate – our sample size needs to be bigger
2. Control the groups better – ours were somewhat serendipitous.
3. Can we try a “fake bad” group? IM should be equally high but with the opposite intent.
4. Explore cases where IM appears to be very high.
5. Compare the “slow response distributions” with the “fast response distributions” at the other end of the Latency Analysis. Is there something about the differences between the two distributions?

But above all we should continue to develop these processes. Technology now allows us to take data and turn it into meaningful information.

Thank you.

---

<sup>i</sup> Bartram, D, "*The Personality of UK managers: 16PF norms for short listed applicants*", J. Occ. And Org. Psych, 1992, **65**,159-172

<sup>ii</sup> Cattell, RB, Eber, H, Tatsuoka, MM, "*Handbook for the Sixteen Personality Factor Questionnaire (16 PF)*", IPAT, 1988