Test-Retest Reliability and The Birkman Method®

Frank R. Larkey & Jennifer L. Knight, 2002

Consultants, HR professionals, and decision makers often are asked an important question by the client concerning their Birkman® report: "How reliable are my results?" The purpose of the current paper is to directly answer this question using common, non-statistical language. A two-week test retest reliability study found that the average reliability scores of 0.81 (Usual), 0.73 (Need) and 0.89 (Interests) were quite high and therefore support the statement that the Birkman Method® is a very reliable instrument for personality assessment.

Was does the term reliability mean? We must first build a conceptual bridge between the question asked by the individual (i.e. are my scores reliable) and how reliability is measured scientifically. This bridge is not as simple as it may first appear. When a person thinks of reliability, many things may come to mind – my friend is very reliable, my car is very reliable, my internet bill-paying process is very reliable, my client's performance is very reliable, and so forth. The characteristics being implied cluster around concepts such as consistency, dependability, predictability, variability and other related terms. Note that implicit in making these "reliability statements" is the observation that people's behavior, machine performance, data processes, and work performance may sometimes not be reliable. While this may seem quite obvious, the first important concept is that reliability, by definition, will vary over time. The question is "how much does the characteristic of interest vary over different observations?"

How do we know if "something" is reliable? In short, we must measure it. This leads to a second important concept. Typically, when making a judgment that a friend, car, process, or performance is reliable, we use some means of measurement. It is often fairly subjective (e.g., my friend is always there when I need him/her, my car always starts the first time, I have never had an error in internet bill-paying, or my employee is always in the office on time). In this case you are using your cognitive abilities as your measuring device. If you carefully recall and scrutinize your past judgment of the reliability of people, machines, processes, and behaviors you would probably recall an incident or two when your car did not start the first time or your employee may have been late in getting to work. The fact is that you are either categorizing this kind of behavior as non-typical and thus irrelevant or you may be doubting your own objectively in the matter - maybe you did not turn the key far enough or perhaps the employee was on time and you just did not notice. The question in this case is "could the variance in this characteristic be due to a limitation in my ability to measure the characteristic perfectly?"

Do other factors affect reliability? In the examples mentioned above, you may also attribute inconsistent behavior to "context" factors. For example, the car always starts except when the temperature is below 10 degrees, or the employee is probably late because of an unexpected traffic jam. Whatever the behavior or process we are measuring, most people often attribute non-typical behavior to factors outside the control of the individual (e.g., "clearly Jane must have been under a lot of stress to make that big of a mistake!"). Errors are often caused by a change in the environment that is completely independent of individual inconsistency and variance in the measurement tool.

How do we determine if inconsistency is due to the person, the measurement instrument, or the context? Let's pull together the three topics discussed above. The reliability/unreliability of Birkman® scores can be attributed to: 1) individual differences in responses between the first and second assessment session; 2) Birkman® items not producing consistent results; 3) environmental factors during two separate sessions being so different as to significantly affect individuals' responses. Scientists measure the effects of these factors on reliability through application of the scientific method, and in the present analysis applied a test-retest design. The results for the latest Birkman® test-retest study are shown in the table below.

Two Weeks Test-Retest 2002

Components	Usual	Need	Interests	
Esteem	0.81	0.70	Persuasive	0.93
Acceptance	0.85	0.76	SocialService	0.88
Structure	0.75	0.72	Scientific	0.92
Authority	0.82	0.59	Mechanical	0.96
Advantage	0.80	0.77	Outdoor	0.94
Activity	0.88	0.72	Numerical	0.89
Challenge	0.72	0.72	Clerical	0.86
Empathy	0.88	0.78	Artistic	0.89
Change	0.80	0.74	Literary	0.90
Freedom	0.77	0.78	Musical	0.72
Thought	0.78	0.77	ave	0.89
ave	0.81	0.73	stdev	0.07
stdev	0.05	0.05		Range
	Range	Range		.72 to .96
	.72 to .88	.59 to .78		

All significant at the .0001 level

The goal of the rest of the paper is to explain in common terms how to interpret the data represented in the current sample of 77 employees at a large petrochemical organization.

.

Unreliability Attributed to the Individual

Reliability is a group rather than individual measure. Imagine that you are taking a multiple-choice exam in Management 101. You have always been an “A” student. Your exam is returned to you with a score of 82. How reliable is your score? If you took the test over immediately would your score be exactly the same? Besides the knowledge that you gained to prepare for the exam and given the fact that during a second testing you might not remember as much information, what other factors affect the “reliability” of your scores? Perhaps, the first exam was given in the morning and you did not have the time to have your second cup of coffee. You might argue that you missed a few questions because you were not yet awake. Or you could argue the opposite, that you had four cups of coffee and subsequently had so much caffeine that you could not keep your attention focused. In either case, you might tell yourself that 82 does not accurately measure your knowledge since your average score in the class is a 96. In a word, your individual response has changed due to an internal change.

Indeed, clients talk about reliability typically only in regards to their own individual scores. However, the scientific method is applied not to an individual's score but to a group of scores. So when instrument reliability is discussed the unit of analysis is not the individual but a group. Therefore, an instrument might demonstrate a very high reliability but some individual scores will vary greatly. One way of thinking about this is that most people would be able to predict that bubbles would start appearing in a pan (boiling) on a range at 212°F. However, it is virtually impossible to predict where and which individual bubble would be first. While there are exceptions, science typically measures groups and not individuals. In any case, the measure of reliability reflects differences averaged across groups. Individual unreliability is greatly reduced by sampling a group of individuals. We can see the application of this principle by referring to the data below selected from Table 1 (see page 3).

Components	**Usual**	**Need**
Esteem	**0.81**	**0.70**

The numbers (statistics) for Esteem Usual and Need reflect the strength of the relationship between responses in the first session and responses in the second session (this statistic is a correlation and is often referred to as the *coefficient of stability*). The important concept here to understand is that these scores reflect the average of the products (standardized) of all 77 people participating in the study. Individual differences between the two sessions are not singled out for analysis but measured as part of the group. If all participants would have made exactly the same responses on all items, the relationship between the two group scores would have been 1.00 – this would mean a perfect one-to-one relationship on each item for all members of the study. If all participants answered exactly the opposite on each item, the relationship between the two sessions would have been –1.00. If the participants secretly conspired to answer all items randomly, the relationship would have been fairly close to 0.00.

In general, correlations of .70 and above are considered highly related and correlations of .60 and even .50 often are reported in personality assessment reliability studies. A simulation exercise is provided in Appendix 1 so that the reader can actually manipulate the data to learn how the correlation can change depending on the numbers used.

To summarize, individual differences between the two scores are expected and serve as the basis for computing the average strength of the relationship on each component.

Unreliability Due to the Instrument

The purpose of a test-retest design is to focus on the reliability of the instrument. In the current study, 200 employees from a major petrochemical service company were invited to participate in a scientific study jointly sponsored by the company and Birkman International. Anonymity was guaranteed and employees were told that they would be required to take two assessments. Anonymity was guaranteed because previous research has demonstrated that if employees feel that their results may be used in any evaluative sense the answers are biased towards socially desirable responses. Although they were told that a second assessment would take place exactly two weeks after the first, no mention was made that they would take the Birkman® twice. The purpose of this "design" is to limit individual differences in scores due to knowledge that they would be retaking the Birkman® questionnaire . Again, research has demonstrated that knowledge of the design or feedback between sessions biases participants' responses.

Table 1 on page three displays both the individual Component and Interest reliability scores as well as the average reliability across all interests. These average reliability scores of .81 (Usual), .73 (Need) and .89 (Interests) are high and therefore support our earlier statement that the Birkman Method® is reliable.

Unreliability Due to Environmental Factors

There are many environmental factors that can influence individual responses. Some examples are distractions such as noise and interruptions, differing time constraints, workplace and personal stress, and major life changes between testing periods. Unreliability due to environmental changes were minimized by providing clear instructions, having all employees take the Birkman® at their personal computers, having strong managerial support for experimental control, and measuring individual time used for assessment.

In summary, any scientific study assumes that some error will enter into the analysis due to factors beyond the control of the researchers. The positive results of the analysis support the efforts made by the researchers and the design of the study to control for factors outside the control of the employees.

Summary of Test-Retest Results 2002

Table 1 provides the strength of relationship between tests for each Component and Interest score. In addition, the average of all Components and Interests scores are reported so that the reader can compare these results to other instruments. The standard deviation scores are reported to indicate the variance of the score. Finally, the range of Usual, Need, and Interest scores are presented as a second measure of variance.

A History of Reliability. Roger Birkman and his colleagues have historically considered the reliability and validity of The Birkman Method® to be absolutely essential. On the following page a table is provided that reflects test-retest results over many years. Table 2 shows the history of the high reliability of the Birkman®.

Table 2

	Immediate: N = 42		Two-Weeks: N = 132		15-Months: N = 50		Two-Weeks 2002: N = 77	
	Usual	Need	Usual	Need	Usual	Need	Usual	Need
Esteem	0.78	0.87	0.84	0.80	0.32	0.64	0.81	0.70
Acceptance	0.79	0.94	0.80	0.85	0.21	0.40	0.85	0.76
Structure	0.78	0.77	0.70	0.71	0.48	0.57	0.75	0.72
Authority	0.79	0.76	0.74	0.70	0.24	0.50	0.82	0.59
Advantage	0.87	0.83	0.81	0.80	0.60	0.69	0.80	0.77
Activity	0.69	0.86	0.52	0.48	0.48	0.45	0.88	0.72
Challenge							0.72	0.72
Empathy	0.82	0.91	0.84	0.82	0.56	0.45	0.88	0.78
Change	0.81	0.82	0.77	0.76	0.62	0.44	0.80	0.74
Freedom	0.71	0.76	0.81	0.80	0.40	0.60	0.77	0.78
Thought	0.71	0.78	0.74	0.76	0.55	0.62	0.78	0.77
ave	**0.78**	**0.83**	**0.76**	**0.75**	**0.45**	**0.54**	**0.81**	**0.73**
stdev	**0.06**	**0.06**	**0.10**	**0.11**	**0.15**	**0.10**	**0.05**	**0.05**
	R=.69to.87	R=.76to.91	R=.52to.84	R=.48to.85	R=.21to.62	R=.40to.69	R=.72to.88	R=.59to.78
Persuasive			0.72		0.65		0.93	
SocialService			0.81		0.75		0.88	
Scientific			0.70		0.81		0.92	
Mechanical			0.80		0.79		0.96	
Outdoor			0.75		0.72		0.94	
Numerical			0.58		0.73		0.89	
Clerical			0.84		0.78		0.86	
Artistic			0.81		0.81		0.89	
Literary			0.76		0.76		0.90	
Musical			0.81		0.81		0.72	
ave			**0.76**		**0.76**		**0.89**	
stdev			**0.08**		**0.05**		**0.07**	

Table 2 was taken from The Birkman Method ® Reliability and Validity, 2002 (Please see the entire report for additional reliability measures such as Cronbach's alpha which measures the internal consistency of a scale).

The Birkman Method® test-retest data have been collected over twenty years, and the strong research tradition here continues to drive the Birkman®. These research efforts have continually revealed the impressive reliability of the Birkman Method®, and they exemplify the historical and current commitment of Birkman International to research integrity.

Appendix 1 – Simulation of Reliability Statistic

Use the spreadsheet above to learn the relationship between individual scores and the coefficient of stability.

	A	B	C	D	E
1			**Session 1**	**Session 2**	
2		**Employee**	**Esteem 1**	**Esteem 2**	
3		**1**	**1**	**1**	
4		**2**	**10**	**10**	
5		**3**	**20**	**20**	
6		**4**	**30**	**30**	
7		**5**	**40**	**40**	
8		**6**	**50**	**50**	
9		**7**	**60**	**60**	
10		**8**	**70**	**70**	
11		**9**	**80**	**80**	
12		**10**	**90**	**90**	
13		**11**	**99**	**99**	
14					
15		**Correlation**		**1**	
16		Formula is =CORREL(C3:C13,D3:D13)			
17					
18			**Session 1**	**Session 2**	
19		**Employee**	**Esteem 1**	**Esteem 2**	
20		**1**	**1**	**99**	
21		**2**	**10**	**90**	
22		**3**	**20**	**80**	
23		**4**	**30**	**70**	
24		**5**	**40**	**60**	
25		**6**	**50**	**50**	
26		**7**	**60**	**40**	
27		**8**	**70**	**30**	
28		**9**	**80**	**20**	
29		**10**	**90**	**10**	
30		**11**	**99**	**1**	
31					
32		**Correlation**		**-1**	
33		Formula is =CORREL(C20:C30,D20:D30)			
34					
35			**Session 1**	**Session 2**	
36		**Employee**	**Esteem 1**	**Esteem 2**	
37		**1**	**1**	**60**	
38		**2**	**99**	**50**	
39		**3**	**40**	**80**	
40		**4**	**50**	**99**	
41		**5**	**80**	**1**	
42		**6**	**70**	**40**	
43		**7**	**90**	**20**	
44		**8**	**20**	**10**	
45		**9**	**80**	**90**	
46		**10**	**10**	**30**	
47		**11**	**30**	**70**	
48					
49		**Correlation**		**-0.076942**	
50		Formula is =CORREL(C37:C47,D37:D47)			

Duplicate these simulation scores for 11 employees. Once you have entered the scores, use the formula at the bottom of the column to generate the coefficient of stability (correlation). Then change scores to simulated changes in the individual scores to see how much these changes impact the correlation for reliability as displayed in the Table.

This table displays a case where each employee answers exactly the opposite in session 2 as compared to session 1.

This table displays a case where each employee answers randomly during both session 1 and 2.

Statistical Definitions

Reliability: The degree to which a score is stable and consistent when measured at different times (test-retest reliability), in different ways (parallel-forms and alternate-forms), or with different items within the same scale (internal consistency).

Correlation: An assessment of the degree of linear relationship between two continuous variables. Correlation coefficients can range from –1.00 (a perfectly negative relationship to +1.00 (a perfectly positive relationship).

Test of Significance: A test used to determine, in this case, if two variables are related to one another (which would be a statistically significant correlation) or if they are unrelated to one another (which would be confirmation of the "null hypothesis"). A correlation is said to be statistically significant if the probability of the null hypothesis being true (i.e., that the two variables are unrelated) is less than 5 times in 100, or .05.

Cronbach's Alpha: A statistical measure of internal consistency which measures how well items on a scale "fit" together and measure the same construct.

Validity: The extent to which a test measures what it purports to measure. This is often not a simple "yes or no" answer, as there are many types of validity that need to be assessed (i.e., construct, content, criterion-related validity) when developing a personality assessment.

Recommended References

For more information about test-retest reliability, please consult the following sources:

Aamodt, M. G. (1991). Applied industrial/organizational psychology. Belmont, CA: Wadsworth Publishing Company.

Anastasi, A. (1982). Psychological testing. New York: Macmillan.

Birkman, R. W. (2001). The Birkman Method ® Reliability and Validity Study. Houston, TX: Birkman International, Inc.

Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996). Psychological testing and measurement: An introduction to tests and measurement. Mountain View, CA: Mayfield Publishing Company.

Pinneau, S. R. (1961). Changes in intelligence quotients from infancy to maturity. Boston: Houghton Mifflin.