

Scientific Evidence Shows That an Individual's First Set of Responses on the Caliper Assessment is Likely to be the Most Reliable Set of Responses

ABSTRACT

The consistency of scores on an assessment from one time to another is another way of referring to the stability, or reliability, of the results from the assessment. A number of factors can affect the reliability of scores from one test session to another. These factors can lower the degree to which scores obtained on an assessment at time 1 are consistent with those gathered on the same assessment at a later time (time 2). The nature of these factors is presented in the body of this document. With these factors in mind, Caliper has conducted scientific studies on three different samples of individuals to address the stability of scores on the Caliper assessments. All of these studies were conducted using a scientific method called a "test-retest research design." The average stability of the Caliper results across these three samples of working adults was quite high (average stability coefficient = +0.81), indicating that the results gathered from administering the first Caliper Profile (time 1) are very stable and reliable.

PROPERTIES OF THE CALIPER ASSESSMENT

Caliper has used the scientific test-retest research design to assess the reliability or stability of the Caliper measurements. This research method requires the administration of the exact same assessment, and its items, in exactly the same order, to the same sample of individuals, across a specified time period. A typical test-retest time interval is two weeks. It is desirable to conduct the study on individuals who have the characteristics of the population to whom the findings are being compared. Caliper has conducted three scientific studies within the past four years to assess reliability or stability of the Caliper assessment. Although Caliper conducted two studies with two-week testing intervals, our most recent study involved a one-year testing interval. The average correlation coefficient across the three studies is +0.81, indicating that the traits measured by the Caliper assessment are highly reliable, meaning that the scores obtained during the first administration (time 1), are highly likely to be obtained during the second administration (time 2).

ENVIRONMENTAL CONDITIONS AT THE TIME OF ADMINISTRATION OF THE CALIPER ASSESSMENT

Environmental conditions can refer to such factors as instructions provided to the examinee, the amount of time provided to complete the assessment, noise, interruptions, workplace stress, whether

the assessment was monitored by a proctor, or whether the assessment was administered remotely or at the employer's site. Any of these factors can influence the reliability of results, by producing changes in an examinee's responses from time 1 to time 2. In order to obtain accurate estimates of reliability of results, it is recommended that the level and effects of these factors are not only minimized, but are also kept essentially the same for both test administrations.

SCORING THE CALIPER ASSESSMENT

With some assessments, it is possible that error in measurement, which has the potential to reduce the reliability of the results of the assessment, can occur due to differences in scoring procedures from time 1 to time 2. This can be particularly true for assessments that are subjectively scored. However, Caliper assessments are objectively scored, and thus measurement error due to differences in scoring procedures is highly unlikely with respect to the Caliper assessments. Our objective scoring programs have been verified and tested across numerous conditions. Furthermore, the responses on the Caliper assessments are normed, or thus interpreted, by converting the raw scores to percentiles derived from a normative database of Caliper scores of individuals in the same country/language as the examinee. Once verified and tested, these objective scoring programs become part of our automated scoring procedure, assuring that responses on the Caliper assessments are scored (and normed) the exact same way, each and every time.

INTERPRETING THE RESULTS ON THE CALIPER ASSESSMENT

As mentioned above, for any given individual, responses (raw scores) on the Caliper assessments are converted to percentiles derived from a normative database of Caliper results from others similar to the examinee in country/language. Attempts to interpret an individual's results using another country's norms will, because of cultural differences in personality trait scores, yield differences in interpretation from time 1 to time 2, and thus the results of the two assessments will appear dissimilar.

Another type of error that can reduce the reliability or stability of results from time 1 to time 2 can be attributed to the individuals who interpret or "rate" the

percentile results. An index of this source of measurement can be obtained by presenting two different groups of raters with the same set of assessment results. One would then calculate the degree of agreement among the two groups of raters; this index is called "inter-rater reliability." Caliper has conducted three studies to investigate inter-rater reliability. We are pleased to report that there is an extremely high degree of agreement between raters who interpret the Caliper results; the average inter-rater reliability is 85%.

EFFECTS OF FEEDBACK FROM TIME 1 RESULTS ON TIME 2 RESPONSES

Individuals often receive feedback, or even seek feedback formally or informally, after they have completed an assessment (time 1). Published literature indicates that regardless of the source of the feedback, feedback after time 1 can have significant effects on responses gathered from the same individuals on the exact same assessment at time 2, particularly if feedback is received shortly after time 1. Sometimes the feedback

Published literature indicates that regardless of the source of the feedback, feedback after time 1 can have significant effects on responses gathered from the same individuals on the exact same assessment at time 2...

involves an analysis of the candidate's specific results on the assessment. Oftentimes the feedback is more subtle. For example, if the assessment is given in an employee selection context, getting hired, as well as not getting hired, is critical feedback to the candidate. Likewise, being asked to complete the assessment for a second time can signal to the employee that the employer is not satisfied, in some way, with the employee's responses or performance, and can cause the employee to change his or her true responses on a second administration of the assessment.

Two separate Caliper studies conducted in the 1980s illustrate the effects of time 1 assessment feedback on the reliability estimates of time 1/time 2 results. Individuals in these samples were job applicants at time 1; however, they were job incumbents (employees in position) at time 2. The average correlation between time 1 and time 2 results was quite low (stability coefficient = +0.56). When this correlation is compared to the correlation obtained by

Caliper following the use of the test-retest scientific method previously discussed (+0.81), it is obvious that feedback to the job incumbents (either direct or indirect) played a role in lowering the correlation between these two sets of assessment results.

This finding also has important implications for the meaning, or usefulness of the results to the employer, as the maximum possible validity of an assessment, with validity being defined as the correlation between the results of the assessment and some independent measure of (job) performance, is equal to the square root of the stability coefficient. Thus, the lower the reliability, or stability of the scores on the assessment, the lower the relationship between the assessment results and job performance.

INDIVIDUAL DIFFERENCES

Most all scientific, statistical studies are based upon measuring groups of individuals, not single individuals. This is true of studies that estimate the reliability, or stability of scores on personality or cognitive

assessments. All estimates of reliability or stability of a personality or cognitive assessment are based upon the average results of a group of individuals. This means that even though the stability coefficient of the Caliper assessments is highly reliable (+0.81), this coefficient was calculated by averaging the results of a group of individuals. Note that there is not a perfect or 1:1 relationship between scores obtained at time 1 and those at time 2. If this was the case, the stability coefficient would be equal to 1.00. Instead, a coefficient of 0.81 indicates that for some individuals, on some traits, scores vary from time 1 to time 2. Furthermore, due to individual differences, scores of some individuals are likely to vary more than others from time 1 to time 2. Overall, however, evidence presented in this document fully supports the claim that for the majority of candidates who are administered the Caliper assessment, results obtained at time 1 are highly consistent with those obtained at time 2. ■

RECOMMENDED REFERENCES

Anastasi, A. (1982). Psychological testing. NY: Macmillan.

Burger, J.M. (2005). Personality (6th edition). Belmont, CA: Wadsworth/Thomson Learning.

Caliper Technical Manual (4th edition). (2005). Princeton, NJ: Caliper, Inc.

Dihoff, R.E., Brosvic, G.M., & Epstein, M.L. (2003). The role of feedback during academic testing: The delay retention effect revisited. Psychological Record, 53, 533-548.

Hausknecht, J.P., Trevor, C.O., & Farr, J.L. Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. Journal of Applied Psychology, 87, 243-254.

Heatherton, T.F., & Weinberger, J.L.. Eds. (1994). Can personality change? Washington, DC: American Psychological Association.
Helmstadter, G.C. (1964). Principles of psychological measurement. NY: Appleton-Century-Crofts.

Lievens, F., Buyse, T., & Sackett, P. (in press). Retest effects in operational selection settings: Development and test of a framework. Personnel Psychology.

Muchinsky, P.M. (2003). Psychology applied to work: An introduction to industrial and organizational psychology (7th edition). Belmont, CA: Wadsworth/Thomson Learning.

Pervin, L.A., Cervone, D., & John, O.P. (2005). Personality: Theory and research (9th edition). John Wiley & Sons, Inc.

Nunnally, J.C., & Bernstein, I.H. (1994). Psychometric theory (3rd edition). NY: McGraw-Hill, Inc.

Schinkel, S., van Dierendonck, D., & Anderson, N. (2004). The impact of selection encounters on applicants: An experimental study into feedback effects after a negative selection decision. International Journal of Selection & Assessment, 12, 197-205.