

Korrespondens mellan David Bartram och Lennart Sjöberg om EFPA-kriterier för testgranskning

Dear professor Bartram,

I have a few questions about the EFPA guidelines for evaluation of psychological tests, used here in Sweden.

1. The guidelines state: "guidelines on sample sizes are based on power analysis of the sample sizes needed to find moderate sized validities if they exist". Apparently, the result is a requirement of $N=100$ which is used as a rigid benchmark, i.e. no validation studies with smaller samples are even considered? How was the power analysis carried out? And why should not a strong correlation in a smaller sample be even considered? This is a very serious question since strict adherence to the sample size requirement makes it extremely difficult, time consuming and expensive to carry out validation studies.

2. What is an "external" criterion? I can see no definition except the phrase "not part of the instrument". Is there a more precise or narrow specification?

3. Construct validation does not mention, as far as I can see, measurement error and the attenuation of correlations due to non-perfect measurements. Should construct validities be corrected for measurement error? If the answer is no, why not?

Yours sincerely,

Lennart Sjöberg

From: Dave Bartram <Dave.Bartram@shlgroup.com>
To: Lennart Sjöberg <lennartsjoberg@gmail.com>
CC: Pat Lindley <patlindley@btinternet.com>
Date: Wed, 9 Jun 2010 13:08:33 +0100
Subject: RE: EFPA Criteria
Thread-Topic: EFPA Criteria
Thread-Index: AcsHu+UgZRg6oP0ASQC6vnGQxlCMjQACwqRw
Accept-Language: en-US, en-GB
X-MS-Has-Attach:
X-MS-TNEF-Correlator:
acceptlanguage: en-US, en-GB

Dear Prof Sjöberg,

Thank you for your email. In relation to your questions I would start by highlighting the Introduction to Section 7:

"It is almost impossible to set clear criteria for rating the technical qualities of an instrument. Under some conditions a reliability of 0.70 is fine; under others it would be inadequate. A criterion-related validity of 0.20 can have considerable utility in some situations, while one of 0.40 might be of little value in others. For these reasons, summary ratings should be based on your judgement and expertise as a reviewer and not simply derived by averaging sets of ratings.

These notes provide some guidance on the sorts of values to associate with inadequate, adequate, good and excellent ratings. However these are intended to act as guides only. The nature of the instrument, its area of application, the quality of the data on which reliability and validity estimates are based, and the types of decisions that it will be used for should all affect the way in which ratings are awarded."

The term 'benchmark' only occurs once in the EFPA document and then in relation to the descriptors given for section 2.2 – not the Evaluation sections.

Thus N=100 is not a rigid benchmark (the documents states "e.g. [for example] sample size less than 100") and smaller n studies are certainly relevant if they find significant higher r values. The basis for the benchmark is Table 3.3 in Cohen's Statistical Power Analysis of the Behavioural Sciences (2nd Edition, 1988). If one sets as a target a correlation of 0.30 (which Cohen classified as a 'medium effect size') then a sample size of n=92 is required for the probability of rejecting the Null Hypothesis being 0.90 or better (i.e. Type 2 error of 0.10 for the test of the hypothesis that r=0 with a Type I error set at 0.05). 90% power is conventionally used as a reasonable benchmark and 0.05 as a maximum type 1 error for significance testing. In short, the guidelines (and they are guidelines not benchmarks) address the 'typical' case where you are looking for moderate correlations. In such cases n=100 is a safe number. From the same table a large correlation (r=0.50) could be detected with power 0.90 and significance level 0.05, with an n=30, while a small effect (r=0.10) would require n=900. For this reason, construct validity studies which are looking for convergent validity evidence (i.e. that comparable scale do correlate) will typically be able to get away with smaller samples sizes (as they are looking for correlations of r=0.5 or more), however, their power to detect evidence of divergence (i.e. that different scales do not correlate) will be very limited if the sample size is small.

'External criterion' is intended to related to measures that lie outside the construct space of the test. Such criteria are typically work-related measures of performance, ratings of competencies, etc. Correlations with scales of other instruments designed to measure similar constructs, on the other hand would be regarded as providing evidence of construct validity. As such they help inform us about the nature of the constructs the test measures, but do not show us what 'external' measures it relates to.

The guidelines provided are for typically uncorrected (i.e. attenuated) measures where reliability is concerned. Clearly, if you correct measures of stability for unreliability, you should get a correlation of 1.00.

I would expect people to correct validity coefficients that have suffered range restriction, as may be the case in selection validation studies. Such coefficients might also be corrected for the reliability of external criterion measures. Thus, we would look at what are generally referred to today as 'operational validities'. Having said that, I would expect test manuals to always report both corrected and uncorrected values.

I hope these answers address your queries.

I have copied this to Pat Lindley for information, as she is the BPS Senior Test Reviews Editor and operates with the EFPA Guidelines.

Best regards

Dave Bartram

Dear Professor Bartram:

Thank your swift response. I am still uncertain about construct validation and correction for attenuation, however. Corrected correlations will always be higher, of course, and hence more likely to meet the requirements.

It seems to me now that I have read your response, that $N=100$ is recommended in order to get a good chance to detect a medium size correlation. This seems very reasonable, but suppose that a study with $N=60$ results in a validity of 0.6? Should the study be ignored because the "requirement" of $N=100$ was not met? Tests are assessed here in Sweden as if $N=100$ is absolutely necessary, otherwise the results are dismissed. In other words, a *recommendation* has been misread as a *requirement*.

A few more questions related to the EFPA guidelines:

1. Construct validation is discussed in terms of correlations but such validation can often be carried out in contexts where a linear correlation is not appropriate. In such cases, would it be acceptable, e.g., to do an ANOVA and estimate η^2 , as a statistic corresponding to r in usual correlation analysis? I have a feeling that such an approach may be ignored since the guidelines are so strongly focused on correlations (which I understand to mean linear Pearson correlations). Another case would be strong curvilinear relationships, resulting in zero linear correlations.

2. Most tests have several subscales. How should one assess validity in such cases? It seems that a common notion is to compute correlations between a criterion scale and each of the subscales, then the median correlation. However, the result could be quite misleading since a multiple correlation, or the correlation with a criterion based on a weighted average of the subscales, could give a very different picture of the value of the test.

Yours sincerely,

Lennart Sjöberg

From: Dave Bartram <Dave.Bartram@shlgroup.com>
To: Lennart Sjöberg <lennartsjoberg@gmail.com>
CC: Pat Lindley <patlindley@btinternet.com>
Date: Thu, 10 Jun 2010 15:54:33 +0100
Subject: RE: Read:
Thread-Topic: Read:
Thread-Index: AcsH0O7BFecSW9JZTjqGk4ePGn48DgA2X+JA
Accept-Language: en-US, en-GB
X-MS-Has-Attach:
X-MS-TNEF-Correlator:
acceptlanguage: en-US, en-GB

Dear Professor Sjöberg,

I'm afraid that if these examples are being used as requirements then the EFPA criteria have been misunderstood. Everything that is in parentheses with "e.g." by it is meant to be illustrative only not prescriptive.

When considering ratings, reviewers should take into account whether they are looking at corrected or uncorrected correlations and bear in mind that uncorrected value underestimate true validity.

As to your questions about construct validity then I would say that any relevant measure can be used – we should not be restricted to correlations. In general one is looking at the magnitude of measures of association – whether Pearson r , eta or anything else – or indeed any other relevant indicator of effect size (e.g. d values).

There is discussion in the notes on how to deal with multi-scale instruments and how to take an overall view of the validity of the instrument from that of the scales. It is always expected that in the review of a test each rating will be accompanied by comments from the reviewers that explain why the particular rating fits the instrument.

In the UK the review process involves two independent reviewers who complete the review process and whose ratings and comments are collated by a consulting editor. The Consulting editor's collation of these are then reviewed by the Senior Editor (Dr Lindley) who can take an overall view on whether ratings have been applied appropriately. At all stages, ratings rely on the knowledge, experience and judgement of the reviewers and not on a literal interpretation of the notes as benchmarks.

I'm sure we agree that any attempt to prescribe specific benchmarks for test quality will fail as so many factors have to be taken into account in coming to a fair and reasoned judgement about the quality of an instrument. The EFPA criteria attempt to provide a standardized framework for reviewing tests but do not purport to set specific quantitative standards for quality.

Your questions are useful in indicating that we may need to do some more work on expanding the explanatory notes and example that go with the criteria. I believe that the STN in Norway is currently reviewing its use of these criteria and we may well take that as an opportunity to pick up some of the issues you have raised.

I hope this is helpful.
Regards
Dave Bartram

Dear professor Bartram,

thank you again for your clarifying and very important answers to my questions. Can I show them to others?

Yours sincerely

Lennart Sjöberg

From: Dave Bartram <Dave.Bartram@shlgroup.com>
To: "lennartsjoberg@gmail.com" <lennartsjoberg@gmail.com>
Date: Fri, 11 Jun 2010 08:24:17 +0100
Subject: Re:
Thread-Index: AcsJNtXJPIIUDnWlSeK1/e6ejt8yXgAAERwo
Accept-Language: en-US, en-GB
X-MS-Has-Attach:
X-MS-TNEF-Correlator:
acceptlanguage: en-US, en-GB

Please feel free to treat these as 'public'. I'm keen to ensure that these issues are widely understood.
Best regards
Dave Bartram