The Construct Behind Content Validity: New Approaches to a Better Understanding

MAIK SPENGLER
S & F Personalpsychologie Managementberatung GmbH

PETRA GELLÉRI AND HEINZ SCHULER
Hohenheim University

Murphy (2009) states that content validity is "neither sufficient nor necessary to predict job performance." At first glance, several studies support his view. Nevertheless, the meaning of the presented data does not seem as clear as suggested; we therefore recommend a re-evaluation of the proposed negligibility of content validity. We first point out that preliminary findings might be the result of a lack of differentiation in interpretation of coefficients without accounting for the nature of included predictors and criteria. Some empirical evidence for the utility of content-valid measures is provided. Finally, we propose two routes, which can aid solution finding in this debate: the usage of relative importance measures and a "multisign–multisample" approach.

According to Wernimont and Campbell (1968), test results can be interpreted as signs or samples. *Signs* refer to measures of latent traits, whereas *samples* represent work behavior in real-life situations. Also,

with regard to content validity, a clear distinction between signs and samples is helpful to understand their unique contributions to predict job performance. Schuler and Schmitt's (1987) theoretical conceptionalization of multimodality advises the use of multiple measures to predict job success (e.g., simulations, tests, and biographical data), as well as preferable approaches to validate those measures. Simulations (samples) are best validated by considering their content, whereas tests (signs) developed to assess traits are checked by the multitrait–multimethod approach (Campbell & Fiske, 1959). The comparison of convergent and discriminant signs provides help in the estimation of construct validity, and biographical data are related to criterion measures.

It is not surprising that in most cases general mental ability (GMA) predicts the ability to perform within a given job, as some signs are generalizable predictors of job performance over a broad range of professions and job levels (Schmidt & Hunter, 1998). This is also true for the personality traits conscientiousness and emotional stability (Barrick, Mount, & Judge, 2001). Because of these findings, one of the main arguments of Murphy's article withholds empirical testing, namely, in most jobs, GMA will explain parts of the variance in applicants' later performances.

Correspondence concerning this article should be addressed to Maik Spengler.
E-mail: spengler@personalpsychologie.de
Address: S & F Personalpsychologie Managementberatung GmbH, Hackländerstraße 17, D-70184 Stuttgart, Germany
Maik Spengler, S & F Personalpsychologie Managementberatung GmbH; Petra Gelléri and Heinz Schuler, Hohenheim University, Chair for Psychology (540F)

Following Brunswik (1955), a major limitation of the interpretation of findings concerning overall job performance is the lack of symmetry in predictors and criteria. Wittmann (1990) points out that such symmetry maximizes their relationship, and he identifies several cases that lead to reduction of correlations. Most important to the current debate is the inappropriate use of a highly specific predictor (e.g., a presentation exercise) as a measure for a general criterion (e.g., overall job performance). Given the small correlations between objective criteria and supervisor ratings (Sackett, Zedeck, & Fogli, 1988), a comparable conceptualization of overall job performance is also doubtful, even more so when multiple studies are aggregated. We will later suggest a research rationale to overcome this limitation.

Findings that indicate incremental validity of content-valid testing. Test hybrids combine the positive aspects of both intelligence assessment and work samples (Klingner & Schuler, 2004). As this test format is based on job analysis, it is specific for a class of jobs. Hence, this relatively new format incorporates construct-oriented (sign) as well as simulation-oriented (samples) aspects. A validation study by Klingner and Schuler reports that a test hybrid shows similar correlations with grades in theoretical examinations when compared with classical tests of GMA but possesses substantial incremental validity in predicting supervisor ratings in the clerical setting it was designed for. This result was even surpassed by an analogous test hybrid for apprentices in technical occupations (Goerlich & Schuler, 2007). Another test hybrid developed and validated by Schuler, Mussel, and Schmidtborn (2008) was specifically designed for an industrial sample of apprentices. The hybrid measure showed higher predictive validity in both criterion measures of supervisory ratings and training success when compared with two different measures of GMA. As a control, clerical apprentices were also assessed with the hybrid test originally designed for technical apprentices. For this sample, the superior predictive validity of hybrid testing diminished and was comparable to those of the two measures of GMA. So far, hybrid tests have shown that a content-related construction of an intelligence test will lead to higher predictive validity in the respective class of jobs when compared with GMA.

Given theoretical similarities and the substantial correlation of test hybrids with the compared measures of GMA, test hybrids obviously measure GMA. Nonetheless, utilizing both aspects of content validity—namely, in both signs (by measuring a relevant construct) and samples (by using job specific content)—is a promising way to enhance predictive test validities. In addition, this finding indicates the importance of content validation because test hybrids have repeatedly shown to be a superior form of a GMA assessment for specific samples. We found a substantial contribution of content-valid predictors in a given set of predictor variables that include general forms of GMA (Schuler et al., 2008). Moreover, the data showed smaller correlations of the test hybrid with personality measures than general forms of GMA, a further indication of the unique contribution of content-valid test formats.

Likewise, the sign and sample approach can be applied to content validity of work samples. Here again, samples highlight job analytically derived facets of a vocation (e.g., presentation skills) in contrast to implicitly measured signs. The latter are broad, generalizable measures of latent traits that enable incumbents to fulfill job requirements, which is the reason for their significant correlations with overall job performance. The incremental validity reported by Roth, Bobko, and McFarland (2005) for work samples over and above intelligence is small, but the authors themselves urge caution because neither predictive validity of GMA nor work samples were corrected for range restriction.

It should further be taken into consideration that content-valid methods are

also often less reliable selection methods, for example, because of insufficient standardization. Furthermore, content-valid job samples seldom embrace all aspects of a given job, so it is obvious that they provide higher predictive validity for specific criteria. Given the fact that work samples do not explicitly measure a single construct (i.e., GMA), it is no surprise that there is a lot of variety in the empirical findings, which can hardly be compared with each other: thus, the ''apples and oranges problem'' is virulent when aggregating work samples in a meta-analytic research paradigm.

We agree with Murphy's argumentation concerning the point that content-valid variables partly show predictive validity because of the fact that they are intercorrelated with signs (e.g., GMA). Yet, the findings presented above indicate that the conclusion that illusory correlations explain the incremental effects of content-valid methods is only a partial solution to the underlying scientific question. In a next step, we would like to provide two different approaches which might shed light on the debate initiated by Murphy.

Two routes to a clearer understanding of content validity. Given that meta-analytic results are not controlled for the effects of intercorrelations described above, we suggest to re-examine the collected data. Murphy outlines the effect of multicollinearity in organizational behavior research when estimating the unique contribution of a single predictor (e.g., a content-valid measure) in a given set of multiple variables that predict a given criterion. Indeed, a given correlation of predictor and criterion is an insufficient measure of predictive validity because this index disregards the influence of further predictors in the analyzed regression model.

As there are specific methods available to take the intercorrelations of multiple predictors into account when measuring the unique effect of a given predictor, researchers are urged to address such a hypothesis empirically, for example, by computing dominance analysis (Budescu, 1993) or by the more convenient relative weights analysis (Johnson, 2000). Both procedures consider the proportionate contribution of each predictor to the criterion (R^2), considering both its correlation with the criterion and its effect when combined with the other variables in the regression equation, and lead to similar results (Johnson). They both estimate the importance of a predictor in the test battery on an empirical basis. We, therefore, suggest future research to test the unique contribution of content-valid test formats in the following design: Instead of reporting incremental validities only, researchers could aggregate relative importance indices in order to obtain results that are controlled for effects of multicollinearity. Such data would allow a clearer view on the unique contribution of content-valid predictors in a given regression model that includes general measures of GMA and personality factors (i.e., Big Five). A meta-analytical approach using relative importance indices (see Whanger, 2002, cited from Johnson & LeBreton, 2004) might also be a fruitful possibility to test Murphy's hypothesis that content validity does not uniquely contribute to the prediction of job success.

As mentioned earlier, the concerns of missing symmetry of predictors and criteria should be more closely examined. In order to make a theoretically based assumption about the relevance of content validity, we suggest the following framework:

First, relevant predictors and criteria have to be identified and specified, meeting the demands for symmetry. Each predictor should be paired with a corresponding criterion. This shall equally apply to overall, composite, and specific measures. Correlations can be analyzed in analogy to Campbell and Fiske (1959) in a ''multisign-multisample matrix'' and examined regarding convergent and discriminant validity. This requires operationalizations of both predictors and criteria classified by the sign and samples approach. For example, possible predictors could be an intelligence

test (GMA predictor), an apprentice test hybrid (hybrid predictor), and a presentation exercise (sample predictor), whereas criteria could be grades (GMA criterion), supervisor's assessment of technical comprehension (hybrid criterion), and a scale from a supervisor rating based on vocational behavior concerning presentations and similar situations (sample criterion). Of course, multiple predictors and criteria for each approach are desirable to enhance reliability.

In a "multisign–multisample matrix," correlations between different predictors of the same category are measures of construct validity. The same applies to measures of criteria. Hybrid predictors should show higher correlations with GMA, as well as with work samples, compared with the correlations between the latter two. In this design, relationships between predictors and criteria for the different measures (both signs and samples) are part of the predictive block. Here, the validity diagonal consisting of predictors and criteria of the same concept of measurement (GMA–GMA, hybrid–hybrid, and sample–sample) can be assumed to show the highest correlation coefficients. Reduced fit between predictor and criterion will lead to lower predictive validity scores, so the lowest correlations should result between sample predictors and GMA criteria and between GMA predictors and sample criteria. Earlier, we have cited some empirical findings concerning test hybrids, which underline the assumed higher correlations, when both sign and sample approach are regarded. Further research is required to compare the assumed rank order of observed correlations in each cell.

In a further step, the suggested analysis needs to be extended to different occupational groups because content validity expects distinct job classes to differ in their specific job requirements. Satterwhite, Fleenor, Braddy, Feldman, and Hoopes (2009) have shown homogeneity of personality and cognitive abilities within occupations. Thus, we assume general measures of GMA to be stable predictors of job

success, whereas content-valid measures most accurately meet the requirements of the population they were designed for and will therefore show less general predictive validity. This means the pattern of correlations in the "multisign-multisample matrix" should hold true only for the target group and will vary for differentiated populations. The test hybrid reported above (Schuler et al., 2008) is one of the empirical findings pointing in this direction.

Although relevant content is neither necessary nor sufficient for predictive validity, it should not prevent researchers from developing content-valid measures, as they nevertheless hold promising additional information about the incumbents knowledge, skills, and abilities (KSAs) and represent a solid base for their later training. Even though incremental validity has been shown repeatedly, its magnitude still remains unclear, as it is methodologically contaminated by several aspects (e.g., multicollinearity or the asymmetry of predictor and criterion measures). We hope this comment will stimulate further research to shed light on this debate.

References

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What we do know and where do we go next? *International Journal of Selection and Assessment, 9,* 9–30.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin, 114,* 542–551.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Goerlich, Y., & Schuler, H. (2007). *Arbeitsprobe zur berufsbezogenen Intelligenz. Technische und handwerkliche tätigkeiten* [A work sample for occupational intelligence. Technical and skilled manual work]. Göttingen: Hogrefe.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research, 35,* 1–19.

Johnson, J. W., & LeBreton, J. M. (2004). History and use of relative importance indices in organizational

research. *Organizational Research Methods, 7*, 238–257.

Klingner, Y., & Schuler, H. (2004). Improving participants' evaluations while maintaining validity by a work sample-intelligence test hybrid. *International Journal of Selection and Assessment, 12*, 120–134.

Murphy, K. R. (2009). Content validation is useful for many things, but validity isn't one of them. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 453–464.

Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58*, 1009–1037.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73*, 482–486.

Satterwhite, R. C., Fleenor, J. W., Braddy, P. W., Feldman, J., & Hoopes, L. (2009). A case for homogeneity of personality at the occupational level. *International Journal of Selection and Assessment, 17*, 154–164.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schuler, H., Mussel, P., & Schmidtborn, A. B. (2008). *Crossing GMA and work samples: Hybrid tests as multimodal conceptualizations.* Poster presented at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco.

Schuler, H., & Schmitt, N. (1987). Multimodale messung in der personalpsychologie [Multimodal measurement in personnel psychology]. *Diagnostica, 33*, 259–271.

Wernimont, P., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372–376.

Wittmann, W. W. (1990). Brunswik-symmetrie und die konzeption der fünf-datenboxen. Ein rahmenkonzept für umfassende evaluationsforschung [Brunswik-symmetry and the conception of the five data-boxes. A framework for comprehensive evaluation research]. *Zeitschrift für Pädagogische Psychologie, 4*, 241–251.