Original Article

Experimental Test Validation

Examining the Path From Test Elements to Test Performance

Stefan Krumm,[1] Joachim Hüffmeier,[2] and Filip Lievens[3]

[1]Institute of Psychology, Freie Universität Berlin, Germany
[2]Institute of Psychology, TU Dortmund University, Germany
[3]Department of Personnel Management, Work & Organizational Psychology, Ghent University, Belgium

Abstract: Although the vast majority of validation studies rely on correlational validity evidence, there is an increased recognition that validity should also focus on whether variations in the focal psychological attribute lead to variations in the measurement outcomes. Therefore, calls have been made that validity evidence should also be gathered through experiments. Existing experimental validation strategies focus on manipulating psychological attributes and their effects on measurement outcomes. In the current manuscript, we present an additional and complementary approach that focuses on manipulating test elements (instead of psychological attributes) that are considered indispensable for test functioning. Examples from personality, situational judgment, emotional intelligence, and reading comprehension domains are presented to illustrate our approach. The presented approach is integrated into existing validation strategies.

Keywords: validity, validation, test construction

In psychological assessment, the validity of test scores plays a central role. Invalid inferences based on test scores can lead to ineffective or even harmful therapies, wrong personnel selection and development decisions, and inadequate educational interventions. Thus, gathering validity evidence of test scores is a conditio sine qua non for every test developer before making a test available for further use. In addition to most test developers seeking to establish validity evidence through correlation matrices organized along the lines of nomological nets, some researchers posited that it is also important to understand how variation in a psychological attribute (used herein as an umbrella phrase for latent states and traits), which the test intends to measure, causally affects test scores and, to this end, call for experimental validation studies (Bornstein, 2011; Borsboom, Mellenbergh, & van Heerden, 2004; Embretson, 1983).

So far, strategies to set up experimental validation strategies are scarce, though. Building on the notion that tests are experiments (with test items as experimental conditions and test takers' responses as dependent variables; for more details see below), we suggest a complementary way of providing experimental validity evidence: To examine how specific *test elements* (e.g., a text in a multiple-choice reading comprehension test item or a description of a situation in a situational judgment test item), which are considered indispensable for the functioning of a test, causally affect test scores. In focusing on manipulating test elements (instead of psychological attributes), we present an additional option to answer recent calls for more experimental strategies of test validation (e.g., Borsboom et al., 2004). Up front, we want to clarify and emphasize that we are not criticizing correlation-based validation strategies. That is also the reason why we end by integrating our approach into other validation strategies and by illustrating how a combination of different approaches might be most fruitful.

Validity and Validation

The debate on how to define validity and where to look for validity evidence has a longstanding tradition, which is aptly referred to as a "long and winding road" (Newton, 2012, p. 5). To describe this long and winding road in very broad strokes (for excellent in-depth discussions, see Kane, 2013; Newton & Shaw, 2014), the concept of validity has evolved from a property of the test itself – that is, validity as the degree to which a test is measuring what it intends to measure – to a property of the interpretation derived from test scores. It has also evolved from a multifaceted to a unitarian conception. Multifaceted conceptions of validity initially distinguished four types (content, predictive, concurrent, and construct; Cronbach & Meehl, 1955), but were soon after reduced to what is today known as the Trinitarian conception of validity (content, criterion-oriented, and construct; APA, AERA, & NCME, 1966). The Trinitarian view was, however, criticized for

DOI: 10.1027/1015-5759/a000393

the imprecise separability of the facets (Landy, 1986) and for offering multiple ways to establish validity evidence – that is, if one way failed researchers were left with two other ways (Guion, 1980). In response to this critique, many researchers as well as the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) adopted the unitary notion of validity as the degree to which evidence and theory support the interpretation of test scores (see also Kane, 2013). Notably, this notion of validity does not come with an a priori set of validation techniques. Rather, gathering validity evidence is essentially viewed as hypothesis testing and general scientific standards for hypothesis testing can be applied (Landy, 1986). According to the *Standards*, *evidential* validity can be established through four sources:

(i) test content,
(ii) item response processes,
(iii) internal structure, and
(iv) relationships with other variables.

A particularly prominent approach to establishing validity evidence through analyzing relationships with other variables (source iv) was first introduced by Cronbach and Meehl (1955). They suggested to analyze the match between empirically derived correlations and the theoretical assumptions (nomological nets) about the link between the focal and other measurements. Today, this validation strategy has been well accepted among researchers and test developers alike and, in fact, can be considered the standard approach to validation (cf. Bornstein, 2011).

This traditional approach is complemented by recent calls to also use experiments to examine whether interpretations of test scores hold. For instance, Borsboom and colleagues (2004) argued that validation should build an a priori established theory as to how the psychological attributes in question *causally* affect test scores – with the causality assumption being best examined through experimental designs (e.g., Bollen, 1989; Loevinger, 1957). As an example of experimental designs in validation research, Bornstein (2011) presented a process-focused model of validity. In this model, validity is defined as "the degree to which respondents can be shown to engage in a predictable set of psychological processes during testing" (p. 536). Similar to Borsboom et al. (2004), Bornstein's main point is that test validation should start by identifying the processes involved during testing (source ii of the *Standards*). He argues that:

> "by using experimental manipulations to alter respondents' psychological processes during testing and assessing the impact of these manipulations on test scores, strong conclusions can be drawn regarding whether or not a test score is actually measuring what it is thought to measure" (p. 536).

Although few would question that validity evidence derived from experiments is a valuable addition to correlational validity evidence and although recommendations exist regarding how to gather such evidence (e.g., Podsakoff, Podsakoff, MacKenzie, & Klinger, 2013), Bornstein (2011) counted that only 9% of validity studies published in top tier journals used experimental procedures. The other 91% of studies represent the common standard to establish validity evidence on the basis of (a network of) correlations (thus tapping into source iv of the *Standards*). Hence, there is a discrepancy between the increased recognition that validation research might use experimental designs (in addition to correlational designs) on the one hand and their actual application on the other hand. This may also be viewed as a discrepancy between the use of sources ii and iv as specified by the *Standards* (see also recent calls for more research on response processes; Ziegler, Booth, & Bensch, 2013). A potential way to alleviate these discrepancies is to increase the variety of experimental validation strategies that target response processes, so that researchers can draw upon these strategies.

Examining the Path From Test Content to Test Performance

Viewed from an experimental lens, psychometric tests constitute a special form of an experiment, in which test items are the experimental conditions (independent variables) and test takers' responses to items serve as dependent variables. In experimental terminology, test administration among a group of test takers is essentially a within-subjects design; that is, every participant is partaking in every experimental condition. As a result of such an experiment, researchers can examine the effect of test items on average responses of a group of test takers. By manipulating the content of test items, one can further examine whether test item content causally affects item responses.

This is where our approach sets in. We argue that we can provide insights into the validity of tests by manipulating the availability of (presumably) essential elements of items. Test items typically consist of elements that *enable* test takers to complete a task as intended. This may be a matrix of objects with one object missing (to enable reasoning), a description of a situation (to enable situational judgment), or a written text (to enable reading comprehension). To further illustrate, a (multiple-choice) reading comprehension test is essentially "destroyed" if no text is presented, because the assumed psychological process can no longer be performed by any test taker. In this case, test

Table 1. Examining the path from test elements to test performance

(1) Identify test item elements that are essential for the targeted psychological process
(2) Create test versions with and without essential item elements and conduct experiment
(3) Interpret outcome
(4) Evaluate generalizability

Follow-up analyses
- Manipulating smaller parts of test elements (e.g., analyses along the lines of item generation theory)
- Process tracing methods (e.g., think-aloud technique, eye-tracking)
- Correlational validation methods
- Manipulating psychological attributes

performance should drop to chance level. In experimental terminology, either presenting or omitting essential elements is a manipulation check. We posit that such checks (e.g., does omitting the text in a reading comprehension test lead to different results as compared to not omitting the text?) may be a basic, but hitherto rarely taken step of the validation process and complements other approaches. In short, instead of focusing on the causal path that runs from psychological attributes to test scores (Borsboom et al., 2004), we suggest a complementary approach that concentrates on the causal path that runs from test elements to test scores.

Omitting core elements of a test should normally result in test performance dropping to chance level or at least dropping substantially below performances in the original test. If either one of these results is observed, the omitted element(s) can be considered essential (or "radical" in terms of item generation theory; Irvine & Kyllonen, 2002; Lievens & Sackett, 2007) for a test and an initial precondition for the test's validity is met. However, if neither one of these results is observed, several conclusions are possible (adapted from Schroeder & Tiffin-Richards, 2014): First, it may be hypothesized that responding to a test in which essential elements are omitted also requires – to some extent – the same psychological attribute(s) as assessed in the complete test (we refer to this as *shared attribute hypothesis* in the remainder of this manuscript). For example, reading response options might also require reading comprehension. Second, it may be assumed that responding to the test version without essential elements requires other psychological attributes, but that these psychological attributes are correlated with the target attribute (*shared underlying variable hypothesis*). For example, reading comprehension and the ability to compare and evaluate response options (in a multiple-choice reading comprehension test without text) might both draw from the same cognitive resources. Finally, if the shared attribute and the shared underlying variable hypotheses are rejected, the interpretation of test scores may simply be invalid (*invalidity hypothesis*). So, follow-up studies are needed to further elucidate why omitting core elements does or does not affect test results.

Table 1 presents the four steps involved in validating test scores according to our approach. First, researchers must identify the elements considered essential (or "radical") for the targeted psychological process. This is essentially a question of content validity (*Standards* source i) and can, for instance, be done by scrutinizing the arguments presented by test authors as to how test items function. Second, test versions with and without elements that were deemed essential are created and administered. Third, results of the experiment are evaluated and, fourth, interpreted with respect to their generalizability to other samples and similar assessment instruments (e.g., other reading comprehension tests). In sum, this approach answers recent calls for conducting experimental validation and provides an additional avenue for examining whether premises underlying tests are actually true (cf. Borsboom & Markus, 2013).

Existing Examples

Reading Comprehension Tests

Albeit generally rare, this approach was sometimes chosen in the domain of reading comprehension (e.g., Katz, Lautenschläger, Blackburn, & Harris, 1990; Rost & Sparfeldt, 2007; Schroeder & Tiffin-Richards, 2014; Sparfeldt, Kimmel, Löwenkamp, Steingräber, & Rost, 2012). These authors found that – across several multiple-choice reading comprehension tests – omitting the text still resulted in correct responses above chance level. Moreover, they employed within-subjects designs and were therefore able to calculate correlations between reading comprehension tests with and without text. Results consistently showed that both versions correlated moderately to highly with each other (e.g., around .60; Rost & Sparfeldt, 2007). These findings stimulated further analyses which revealed that verbal intelligence may account for the convergence between reading comprehension tests with and without text, thereby lending initial support to the shared

underlying variable hypothesis (Schroeder & Tiffin-Richards, 2014). In sum, these experimental validation strategies in the reading comprehension domain have helped in advancing knowledge about their validity and how to further improve such tests (e.g., by creating response options that require less verbal intelligence).

Situational Judgment Tests

Situational Judgment Tests (SJTs) consist of written job-related situations that are typically followed by a set of multiple-choice response options (Motowidlo, Dunnette, & Carter, 1990). SJTs are widely used in personnel assessment and selection and have stimulated a lot of research in the last decade (Whetzel & McDaniel, 2009). The broad consensus has been that SJTs capture situational judgment which relates to context-dependent knowledge. That is, test takers make situational judgments when responding to the situations described by considering situational demands and specifics of a situation (Rockstuhl, Ang, Ng, Lievens, & Van Dyne, 2015). Thus, the situational descriptions are core components of SJTs in that they are key to their assumed functioning.

A study by Krumm et al. (2015) using the herein suggested approach challenged this view of SJTs. These authors administered the most popular SJT (about team knowledge) either with or without situational descriptions. To illustrate, in the condition without situational descriptions, participants were asked to indicate the most effective response (out of four alternatives). Surprisingly, the authors revealed that for between 43 and 71% of the items it did not make a significant difference whether the situational description was included.

The results emphasize that situational descriptions do not seem to unanimously work in the way they were intended to work by SJT developers. Krumm et al. (2015) conducted a follow-up study to answer why. Using a think-aloud technique, these authors revealed that test takers rely on general knowledge about the effectiveness of the behavior presented in multiple-choice response options. Assuming that SJTs generally capture – to some extent – general domain knowledge (Lievens & Motowidlo, 2016; Motowidlo, Crook, Kell, & Naemi, 2009), this finding speaks to the shared attribute hypothesis. Generally, this approach contributed to advancing theory building in the SJT domain. At a practical level, these results question the effortful and cost-intensive process of generating elaborate situational descriptions with the help of subject matter experts.

Picture-Based Motive Tests

As early as 1938, a technique for assessing motives was developed that showed ambiguous pictures to individuals, who were then asked to tell stories about these pictures (Murray, 1938). Since then, such pictures have been used as stimuli in various implicit motive tests (e.g., McClelland, 1985), with either an open-ended or a close-ended response format. Close-ended response formats consist of a set of statements (e.g., "Here, one can easily be rejected by others"), which test takers are asked to indicate as either fitting or not fitting to the picture (Sokolowski, Schmalt, Langens, & Puca, 2000). Regardless of the response format, the underlying assumption is that pictures arouse motives.

In a recent study, Krumm, Schäpers, and Göbel (2016) put this assumption to the test and administered the Multi-Motive Grid (Sokolowski et al., 2000), a picture-based implicit motive test with a close-ended response format. This implicit motive test assesses three motives (achievement, affiliation, and power) in two components each (hope and fear component). Krumm et al. (2015) administered two versions of the Multi-Motive Grid, one with pictures and one without pictures. In the version without pictures, the standard instruction of the Multi-Motive Grid, to imagine social situations, was simply reiterated, but not a single ambiguous picture was presented. Participants were randomly assigned to one of the two experimental conditions. Results indicated that three out of the six test scores were not significantly different across the two conditions. Specifically, scores obtained for the hope and the fear component of achievement, and for the hope component of power were not statistically different across the two experimental conditions. More in-depth analyses also revealed that responses did not differ between experimental conditions for about 50% of the motive-related statements listed underneath each item.

As for SJTs, manipulating test elements revealed that the motive test under investigation did not work as previously assumed, thereby tentatively questioning the underlying theory of such tests. Again, follow-up studies need to clarify whether, for example, response options can also arouse motives and thus the same psychological attribute is measured by both versions of the Multi-Motive Grid (shared attribute hypothesis).

Social Intelligence Tests

Understanding emotions and cognitions of another person is a core aspect of social intelligence (Weis & Süß, 2005). To gauge this ability, performance-based social intelligence tests typically present social situations and ask test takers to judge emotions and cognitions of a focal person (Baumgarten, Süß, & Weis, 2015). This judgment is then compared against expert judgments or the focal person's actual emotions and cognitions in that particular situation. The implicit assumption is that test takers make inferences on the basis of the focal person's behavior. Typically, however, other cues (such as the social context) are also

available and might serve as a viable basis for social understanding.

In a recent study, Baumgarten et al. (2015) examined whether performance in a social intelligence test is actually driven by observing the focal person (i.e., the social cue) or the context information (e.g., the setting of the social situation, non-focal persons). To this end, these authors manipulated the availability of the social cues and the context. They administered three versions of the same test to a student sample. One group worked on a version in which no social cues were presented, another group on a version in which no context information was available. A third group completed the original test including all the information. Results showed that presenting the context only (and no social cues) led to significantly lower test performances in three out of four test subtests. For the auditory subtest, however, the context only condition did not significantly differ from the original test. Overall, the authors conclude that "the [social] cue is the key" (Baumgarten et al., 2015, p. 42).

In this example, the evidence obtained through an experimental validation procedure, that is, that social cues are essential for test performance in all but one subtest, is pivotal in explaining heterogeneity among subtests and may guide subsequent subtest revision. Furthermore, adding an experimental condition in which a (presumably) non-essential element of the test was manipulated helped in investigating alternative hypotheses about test functioning.

Discussion

In response to calls for validation research to use more experimental designs (Bornstein, 2011; Borsboom et al., 2004), we present a novel approach that complements other validation strategies. In particular, our approach focuses on manipulating test elements (instead of psychological attributes) that are considered indispensable for test functioning. Notably, existing approaches in the domains of reading comprehension, situational judgment, implicit motives, and social intelligence have – until now – *not* been framed as examples of the same validation approach. It is thus a contribution of the current paper to provide a conceptual framing for such approaches and, subsequently, to stimulate more research along these lines.

The proposed approach has several advantages. First, it provides additional ways of testing novel conceptions of validity. For instance, Kane's argument-based approach essentially calls for linking test performance to a psychological attribute that is supposedly measured by the test through a set of valid arguments (e.g., Kane, 2013). For this to be achieved, three valid inferences are necessary: the scoring inference (e.g., can test performances be linked to test scores?), the generalization inference (e.g., do scores from one test generalize to other tests with the same measurement intention?), and the extrapolation inference (e.g., can test scores be extrapolated to traits occurring outside test situations?). Our proposed approach can be viewed as particularly apt to add to the scoring inference, because it provides answers to questions such as "Can situations in SJTs effectively capture features of performance associated with different levels of the psychological attribute?" (for a similar phrasing of this question, see Newton & Show, 2014, p. 138). As mentioned above, our approach also adds to examining validity as viewed by Bornstein (2011) and Borsboom et al. (2004). Second, it is relatively easy to conduct. For most tests, it suffices – as a first step – to administer it as is or without core elements (see examples presented above). We recommend conducting analyses along the lines of our approach at an early phase of the validation process (e.g., before conducting experimental validation studies that manipulate psychological attributes), because results obtained from other validation strategies may be affected due to the test containing items for which core elements do not work as intended. Third, it is applicable not only on the test but also on the item level. In addition to selecting items on the basis of factor loadings, item-scale correlations, or item-fit indices in item response theories, items may be selected on the basis of whether their components work in the way they were intended to. Fourth, deeper insights can be gained by comparing items that show substantial differences when core elements are omitted with those that do not show substantial differences. Additionally, applying designs in which all the psychologically meaningful item components are manipulated will also provide unique insights into potential interactions of item components. For instance, some components (e.g., situational descriptions) may only impact test outcomes when other components are given as well (e.g., specifics of the situation presented in response options; cf. Baumgarten et al., 2015). Hence, the current approach can also help toward building more elaborate theories as to when some features are crucial for test performance and when they are not.

Although we propose to examine the causal path that runs from test elements to test scores instead of the causal path from psychological attributes to test scores (Borsboom et al., 2004), our approach is still in line with the assumption of common-cause latent traits. That is, examining *how* individual differences in latent traits result in individual differences in responses to test items, is not in contrast to assuming that individual differences in latent traits *do* result in individual differences in responses to test items. Essentially, our proposed approach is simply adopting a more fine-grained perspective by considering test elements as

being part of how individual differences in latent traits translate into different test performances.

Potential Follow-up Strategies

Manipulating Smaller Parts of Test Elements

In many cases, omitting core elements of a test will lead to test performances dropping to chance level or dropping substantially below the original test. Thus, the omitted element(s) are indeed essential for test functioning and the conducted "manipulation check" will provide the expected result. Researchers may nevertheless be interested to decompose the previously omitted element(s) into more fine-grained features to find out which features or combination of features specifically lead to a drop in test performances and to what extent. Conceptually, this follow-up is in line with the premises underlying item generation theory (Irvine & Kyllonen, 2002; Lievens & Sackett, 2007). Item generation theory posits that test items may be decomposed into several structural features, which can be classified as either radicals or incidentals. According to item generation theory, item features are considered radicals if they contribute to performance (item difficulty). Conversely, features that are surface characteristics of an item and do not contribute to its difficulty are referred to as incidentals.

Process Tracing Methods

The think-aloud technique (i.e., gathering verbal protocols while test takers are working on a test) is apt to provide insights into the mental processes involved in responding to test items. Krumm et al. (2015) conducted verbal protocol analyses to examine how test takers can infer correct solutions in SJT items presented without situations. Through qualitative and quantitative analyses of the verbal protocols, they discovered that test takers relied on their general knowledge about the effectiveness of responses.

Eye-tracking is another way to uncover the processes involved in responding to test items. For instance, studies conducted with the Raven's Matrices Test revealed that test takers differ in their use of two strategies, constructive matching and response elimination. Constructive matching denotes the strategy to construct an optimal solution based on the information in the matrix and then to compare it with response alternatives. Response elimination is, in short, the strategy to eliminate wrong alternatives. Eye-tracking revealed that individuals differ in their preference for strategies (cf. Vigneau, Caissie, & Bors, 2006). Eye-tracking can thus provide insights regarding the extent to which test elements are actually used (i.e., looked at) when responding to test items.

Correlational Methods

We also suggest to follow up on our approach by examining the validity of scores obtained from the test version in which core elements have been omitted through correlations. As mentioned above, the altered version of the test may still measure the focal psychological attribute (shared attribute hypothesis). For instance, reading comprehension tests without text may still assess reading comprehension because written response options are compared in the response process. One might therefore follow up on our approach by comparing correlations of both, the manipulated and the original version of a test with other instruments. Similar and substantial correlations of both versions with other instruments that also assess the focal psychological attribute speak to the shared attribute hypothesis. If correlations with instruments that also assess the focal psychological attribute differ, but correlations with instruments assessing third variables are similar and substantial, this lends support to the shared underlying variable hypothesis.

Manipulating Psychological Attributes

Particularly compelling validity evidence can be obtained when our approach is combined with experimental manipulations of the psychological attribute (Borsboom et al., 2004). That is, 2 (test element: presented vs. omitted) $\times$ 2 (attribute: high vs. low or average) designs can be employed in which (a) presenting or omitting those elements of a test that are considered essential for capturing the psychological attribute is crossed with an intervention (targeting the focal psychological attribute) and a control group. Under the assumption that a test can validly measure the focal psychological attribute, but only does so when core elements are included, one expects a significant interaction effect, such that the intervention group (high on the attribute) yields higher test scores than the control group (low on the attribute), but only when core elements are presented. Conversely, a significant main effect of the attribute factor and an insignificant interaction effect speak to the shared attribute hypothesis. A significant main effect of the test element factor and an insignificant interaction effect speak to the invalidity hypothesis; as does the absence of any significant effects.

Areas for Future Research

Future applications of our approach may include examining the interaction between stimulus and response components of items. The studies using our approach on SJTs and implicit motive tests suggest that the availability of multiple-choice response options may in some cases question the functioning of the stimulus material. Inferring correct

answers from response alternatives instead of inspecting and analyzing the stimulus component of a test, however, may be a threat to test validity *if* response options do not capture the intended construct. In that case, test developers are well advised to ensure that the process of generating answers to test items is not driven by response options. This call applies to virtually all classes of tests consisting of a stimulus component and correct and incorrect multiple-choice response options, such as cognitive ability tests, SJTs, scholastic aptitude tests, conditional reasoning tests, tacit-knowledge test), and social/emotional intelligence tests.

Tests using item frames may also be subject to analyses along the lines of our approach. Studies conducted on the functioning of contextual frames as applied in frame-of-reference personality tests (e.g., Lievens, De Corte, & Schollaert, 2008) confirmed that these frames indeed seem to work as contextual demands. We posit that our approach can also be used to examine effects of other frames, such as temporal frames in questionnaires and contextual frames in ability tests (e.g., as used in problem-solving tests; Greiff, Wüstenberg, & Funke, 2012).

Another area for future research pertains to picture and video content in tests. Several test traditions have developed picture- and video-based variants of its initial forms. Picture and videos are now part of several SJTs, broad personality tests, clinical assessments, motive tests, value surveys, etc. (e.g., Döring, Blauensteiner, Aryus, Drögekamp, & Bilsky, 2010; Lievens & Sackett, 2006). Our literature review, however, revealed that while some studies are available that compare video versus written tests in the domain of SJTs (cf. Lievens & Sackett, 2006) such studies are sparse in other domains. Future research in these domains might adopt our approach by administering test versions with and without videos or pictures to examine their relevance for the validity of test scores.

Limitations

Some limitations should be acknowledged. First, our approach provides evidence that is necessary but not sufficient for establishing the validity of test scores. However, it may serve as an important first step in collecting validity evidence and, together with other approaches, provide important insights as to when and why test elements contribute to the validity of test scores. Second, one may object that our approach to test validation is not applicable to a frequently used form of psychological tests: self-reports that consist of statements one can agree or disagree to (on a Likert-type scale). Indeed, questionnaires may be hard to decompose into several components (but see examples of contextual and temporal frames; e.g., Lievens et al., 2008). Furthermore, one might object that this approach is limited to tests measuring maximum performance, thereby assuming that test takers need to make use of essential test elements in order to present their best possible performance. While this approach was indeed most frequently adopted for tests measuring maximum performance, the example of the picture-based motive test presented above suggests that it can also be applied to tests capturing typical performance – provided that such tests include elements that are essential for the typical performance to occur. However, we acknowledge that our approach may be most suited for tests measuring maximum performance.

Conclusion

Adopting a view of psychometric tests as experiments, the current paper illustrates how a complementary approach to test validation, which manipulates the availability of (presumably) essential elements of test items, can provide valuable insights. Specifically, our approach focuses on how test elements causally lead to test scores as an initial step in a multi-step validation strategy to ensure that these elements work the way they are intended to. Among other advantages, our approach is relatively straightforward to apply and therefore answers recent calls for more experimental validation approaches and for critical assessments of the rationale behind tests (Ziegler & Vautier, 2014).

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

AERA, APA, & NCME. (1966). *Standards for educational and psychological testing*. Washington, DC: APA.

Baumgarten, M., Süß, H. M., & Weis, S. (2015). The cue is the key. *European Journal of Psychological Assessment, 31*, 38–44. doi: 10.1027/1015-5759/a000204

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17*, 303–316. doi: 10.1177/0049124189017003004

Bornstein, R. F. (2011). Toward a process-focused model of test score validity: Improving psychological assessment in science and practice. *Psychological Assessment, 23*, 532–544. doi: 10.1037/a0022402

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement, 50*, 110–114. doi: 10.1111/jedm.12006

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. doi: 10.1037/h0040957

Döring, A. K., Blauensteiner, A., Aryus, K., Drögekamp, L., & Bilsky, W. (2010). Assessing values at an early age: The Picture-Based Value Survey for Children (PBVS-C). *Journal of*

Personality Assessment, 92, 439–448. doi: 10.1080/00223891.2010.497423

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement, 36*, 189–213. doi: 10.1177/0146621612439620

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology, 11*, 385–398. doi: 10.1037/0735-7028.11.3.385

Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Mahwah, NJ: Erlbaum.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. doi: 10.1111/jedm.12000

Katz, S., Lautenschlager, G. J., Blackburn, A. B., & Harris, F. H. (1990). Answering reading comprehension items without passages on the SAT. *Psychological Science, 1*, 122–127.

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? *The Journal of Applied Psychology, 100*, 399–416. doi: 10.1037/a0037674

Krumm, S., Schäpers, P., & Göbel, A. (2016). Motive arousal without pictures? An experimental validation of a hybrid implicit motive test. *Journal of Personality Assessment, 98*, 514–522.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *The American Psychologist, 41*, 1183–1192. doi: 10.1037/0003-066X.41.11.1183

Lievens, F., De Corte, W., & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *The Journal of Applied Psychology, 93*, 268–279. doi: 10.1037/0021-9010.93.2.268

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology, 9*, 3–22. doi: 10.1017/iop.2015.71

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: a comparison in terms of predictive validity. *The Journal of Applied Psychology, 91*, 1181–1188. doi: 10.1037/0021-9010.91.5.1181

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. *The Journal of Applied Psychology, 92*, 1043–1055. doi: 10.1037/0021-9010.92.4.1043

Loevinger, J. (1957). Objective tests as instruments of psychological theory: Monograph supplement 9. *Psychological Reports, 3*, 635–694. doi: 10.2466/pr0.1957.3.3.635

McClelland, D. C. (1985). *Human motivation*. Glenview, IL: Scott, Foresman & Co.

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology, 24*, 281–288. doi: 10.1007/s10869-009-9106-4

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *The Journal of Applied Psychology, 75*, 640–647. doi: 10.1037/0021-9010.75.6.640

Murray, H. A. (1938). *Explorations in personality*. New York, NY: Oxford.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement, 10*, 1–29. doi: 10.1080/15366367.2012.669666

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Cambridge, UK: Sage.

Podsakoff, N. P., Podsakoff, P. M., MacKenzie, S. B., & Klinger, R. L. (2013). Are we really measuring what we say we're measuring? Using video techniques to supplement traditional construct validation procedures. *The Journal of Applied Psychology, 98*, 99–113. doi: 10.1037/a0029570

Rockstuhl, T., Ang, S., Ng, K. Y., Lievens, F., & Van Dyne, L. (2015). Putting judging situations into situational judgment tests: Evidence from intercultural multimedia SJTs. *The Journal of Applied Psychology, 100*, 464–480. doi: 10.1037/a0038098

Rost, D. H., & Sparfeldt, J. R. (2007). Leseverständnis ohne Lesen? Zur Konstruktvalidität von multiple-choice-Leseverständnistestaufgaben [Reading comprehension without reading? Construct validity of multiple-choice reading comprehension tasks]. *Zeitschrift für Pädagogische Psychologie, 21*, 305–314. doi: 10.1024/1010-0652.21.3.305

Schroeder, S., & Tiffin-Richards, S. (2014). Kognitive Verarbeitung von Leseverständnisitems mit und ohne Text [Cognitive processing of reading comprehension items with and without text]. *Zeitschrift Für Pädagogische Psychologie, 28*, 21–30. doi: 10.1024/1010-0652/a000121

Sokolowski, K., Schmalt, H. D., Langens, T. A., & Puca, R. M. (2000). Assessing achievement, affiliation, and power motives all at once: The Multi-Motive Grid (MMG). *Journal of Personality Assessment, 74*, 126–145. doi: 10.1207/S15327752JPA740109

Sparfeldt, J. R., Kimmel, R., Löwenkamp, L., Steingräber, A., & Rost, D. H. (2012). Not read, but nevertheless solved? Three experiments on PIRLS multiple choice reading comprehension test items. *Educational Assessment, 17*, 214–232. doi: 10.1080/10627197.2012.735921

Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence, 34*, 261–272. doi: 10.1016/j.intell.2005.11.003

Weis, S., & Süß, H.-M. (2005). Social intelligence – A review and critical discussion of measurement concepts. In R. Schulze & R. D. Roberts (Eds.), *An international handbook of emotional intelligence* (pp. 203–230). Göttingen, Germany: Hogrefe.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*, 188–202. doi: 10.1016/j.hrmr.2009.03.007

Ziegler, M., Booth, T., & Bensch, D. (2013). Getting entangled in the nomological net: Thoughts on validity and conceptual overlap. *European Journal of Psychological Assessment, 29*, 157–161. doi: 10.1027/1015-5759/a000173

Ziegler, M., & Vautier, S. (2014). A farewell, a welcome, and an unusual exchange. *European Journal of Psychological Assessment, 30*, 81–85. doi: 10.1027/1015-5759/a000203

Received March 23, 2016
Revision received July 25, 2016
Accepted August 2, 2016
Published online March 10, 2017

Stefan Krumm
Institute of Psychology
Freie Universität Berlin
Habelschwerdter Allee 45
14195 Berlin
Germany
stefan.krumm@fu-berlin.de

