Psychology Science, Volume 48, 2006 (3), p. 369-377

The relationship between the ability to identify evaluation criteria and integrity test scores

CORNELIUS J. KÖNIG[1], KLAUS G. MELCHERS, MARTIN KLEINMANN,
GERALD M. RICHTER & UTE-CHRISTINE KLEHE

Abstract

It has been argued that applicants who have the ability to identify what kind of behavior is evaluated positively in a personnel selection situation can use this information to adapt their behavior accordingly. Although this idea has been tested for assessment centers and structured interviews, it has not been studied with regard to integrity tests (or other personality tests). Therefore, this study tested whether candidates' ability to identify evaluation criteria (ATIC) correlates with their integrity test scores. Candidates were tested in an application training setting (N = 92). The results supported the idea that ATIC also plays an important role for integrity tests. New directions for future research are suggested based on this finding.

Key words: personnel selection, integrity test, ability to identify evaluation criteria

[1] Cornelius J. König, Klaus G. Melchers, and Martin Kleinmann, Psychologisches Institut, Universität Zürich, Switzerland; Gerald M. Richter, Novartis Behring, Marburg, Germany; Ute-Christine Klehe, Psychologisches Institut, Universität Zürich, Switzerland, and Universiteit van Amsterdam, the Netherlands.
The research reported in this article was supported by grant Kl 823/6-1 from the German Research Foundation (Deutsche Forschungsgemeinschaft) to Martin Kleinmann. We thank Thomas Hartstein, Dorit Auge, Katja Nicht, Peter Guzzardi, and Torsten Biemann for their help with the data collection, Alice Inauen for scoring the ATIC data, and Sarah Mannion for editorial assistance. Additional we thank Bernd Marcus for providing the IBES.
Correspondence concerning this article should be addressed to Cornelius J. König, Psychologisches Institut, Universität Zürich, Rämistraße 62, CH-8001 Zürich, Switzerland. Electronic mail may be sent to c.koenig@psychologie.unizh.ch.

Integrity tests – tests aimed at assessing honesty – are widely accepted as useful personnel selection tools for the prediction of both productive and counterproductive behaviors (e.g., Ones, Viswesvaran, & Schmidt, 1993). They also explain incremental variance in job performance when used in combination with measures of cognitive ability (Schmidt & Hunter, 1998). Not surprisingly, the use of integrity tests seems to have increased worldwide, both in the United States and in various other countries (cf., Fortmann, Leslie, & Cunningham, 2002; Marcus, Schuler, Quell, & Hümpfner, 2002). Research has not only shown the predictive validity of integrity scales; it has also addressed the multifaceted nature of integrity scales (e.g., Van Iddekinge, Taylor, & Eidson, 2005), the relationship between integrity and the Big Five (particularly with Conscientiousness, Extraversion, and Agreeableness, see Marcus, Funke, & Schuler, 1997), and most recently, the relationship between integrity and psychopathic personality (Connelly, Lilienfeld, & Schmeelk, 2006). However, as yet, research has not explored the relationship between applicants' ability to identify evaluation criteria (ATIC, cf. Kleinmann, 1993) and integrity test scores.

The idea behind ATIC is the following (see, e.g., Kleinmann, 1993): In personnel selection procedures like assessment centers (ACs), it is not revealed in advance to applicants what exactly is evaluated during the selection process. Thus, applicants do not know whether one kind of behavior will be seen more positively or more negatively than a different kind of behavior. However, if applicants are able to identify the criteria that are used for evaluating their performance, they can use this information to adapt their behavior. For example, if an applicant is able to figure out that cooperation is well judged by the assessors in an AC exercise, she can behave accordingly and may decide not to enforce her position by hook or by crook. As a consequence, she will receive higher ratings than if she had not figured this out.

Although Kleinmann (1993) developed his arguments only with regard to ACs, later research has extended these ideas to other nontransparent personnel selection procedures (e.g., Melchers, Kleinmann, Richter, König, & Klehe, 2004). Evaluation criteria are unknown to applicants not only in ACs, but also in interviews. Applicants in an interview also have to discern what the interviewer is attempting to measure with a particular interview question in order to give an answer that is relevant to the specific dimension evaluated. Finally, whether, in a personality inventory (such as an integrity test), applicants discern the targeted dimension or not should also be important because if they do discern it they can activate information or memories of past behavior that are relevant for the given construct. Indirect evidence for this suggestion comes from prior research that has revealed that grouping together items that belong to the same scale (in contrast to presenting items from different scales in an intermixed manner) leads to a shift in the scale means in the direction of the positive endpoint of the scale (e.g., Franke, 1997). One of several explanations (cf., McFarland, Ryan, & Ellis, 2002) put forward to explain this effect is that grouping together items that belong to the same dimension enhances their transparency and also leads to stronger activation of dimension-relevant information.

Empirically, evidence was found that ATIC correlates positively with AC scores in a study using an application training as the setting (Kleinmann, 1993). This finding was replicated in a recent field study by Preckel and Schüpbach (2005). Furthermore, there is also evidence that ATIC is positively related to interview scores (Melchers et al., 2004). Although this is only correlational evidence that cannot show the assumed causal relationship (i.e., high ATIC leading to high scores), Kleinmann and colleagues (Kleinmann, Kuptsch, &

Köller, 1996; König, Klehe, Richter, Kleinmann, & Melchers, 2005) have developed a design that allows the effects of ATIC to be tested experimentally by manipulating the transparency of the evaluation criteria. If the criteria are made transparent prior to an AC task or an interview, candidates do not have to figure out what recruiters are looking for and it becomes irrelevant how high their ATIC is. Thus, nontransparent personnel selection procedures should become easier and candidates in the transparent group should perform better than candidates in the control group (i.e., the nontransparent condition). This is what has been found for ACs (Kleinmann et al., 1996) and interviews (König et al., 2005), meaning that the causal link does indeed go from ATIC to performance in personnel selection procedures and not vice versa.

In conclusion, ATIC has emerged as an important variable for personnel selection but has only been empirically studied with regard to ACs (Kleinmann, 1993; Kleinmann et al., 1996; Preckel & Schüpbach, 2005) and interviews (König et al., 2005; Melchers et al., 2004), and not with regard to integrity tests as a special form of personality tests. Thus, the aim of this study was to test whether applicants' ATIC in an integrity test correlates with their scores in the integrity test.

Method

Setting and participants

A two-day application training program organized by a German university and a regional branch of the German Bureau of Labor Exchange was attended by 92 participants (45 males, 47 females). The application training offered the chance to take part in different personnel selection procedures and included an integrity test among other selection procedures. Participants were either recent or prospective university graduates who were currently applying for jobs or would be doing so in the near future. On average, participants had been attending university for 4.81 years (SD = 1.93). Twenty-seven percent of the participants had already obtained the German equivalent of a Master's degree. Forty-seven percent of them reported prior work experience. Participants were on average 26.8 years old (SD = 2.83, range between 21 and 36 years). Many participants had a background in business or economics (43.5%) or in natural science (19.6%). Participants had to pay a small fee for the training in order to cover part of the costs and to ensure their commitment. They did not receive any information concerning the objectives of the study.

Integrity test

Marcus (2006; see also Marcus et al., 2002) developed a German integrity test called "Inventar berufsbezogener Einstellungen und Selbsteinschätzungen" (IBES; "job-related attitudes and self-evaluations inventory"), which was used for this study. The test allows for a differentiation between overt and personality-based subscales – two categories that are often used to classify integrity (sub-)tests or items (cf. Sackett, Burris, & Callahan, 1989). The overt integrity subscales are aimed at measuring supporting attitudes towards dishonest behaviors. A sample item is "Do you believe that a person who steals goods from your com-

pany on several occasions should get a second chance?" Personality-based items are aimed at measuring personality facets that are known to be related to counterproductivity (e.g., a lack of conscientiousness). A sample item is "I sometimes procrastinate on important projects if I do not feel like doing them".

Three personality-based subscales (Modesty, Conflict Avoidance, and Sensation Seeking) were excluded from the FES because they seemed problematic in the context of the present study: In the FES, high scores in these subscales are taken as indicators of integrity, even though high scores on items with similar content in typical personality tests would be taken as a lack of a positive trait. For example, the item "I find talking easier than listening" of the Modesty subscale is reverse-coded with regard to integrity, and applicants should therefore not endorse it. Such an item is, however, often used in personality inventories to measure extraversion, which is often a positive predictor of performance (e.g., Barrick & Mount, 1991). All remaining 94 items (60 items from 4 overt subscales and 34 items from 2 personality-based subscales) were scored on a 5-point Likert-type scale.

ATIC

The way in which ATIC was measured mirrored the procedure that Kleinmann (1993) developed for assessing ATIC in an AC. The general idea in Kleinmann's study was to assess whether participants had figured out which dimensions were assessed in a particular AC task. Thus, participants were asked to write down the hypotheses that they had entertained while working on an AC task, and later on they were requested to indicate which of their hypotheses corresponded to which of several possible AC dimensions.

Even though participants wrote down their hypotheses concerning all of the AC exercises employed in Kleinmann's (1993) study, writing down hypotheses for all 94 integrity test items would impose an enormous burden on the participants, making an adaptation necessary for the present investigation. We therefore decided to present only some items from each subscale to our participants and use the degree to which they were able to figure out the correct dimension for those items as a measure of their overall ATIC. Accordingly, we created item triplets for each of the six subscales. The first selection criterion for these item triplets was that items had to be phrased positively with regard to the positive end-point of the respective subscale names (i.e., we excluded all reverse-coded items). One of the authors then judged which three of the remaining items were most characteristic for the subscale.

After filling out the integrity test, participants had to write down their hypotheses for each of the triplets. The ATIC questionnaire provided space for a maximum of two hypotheses for each triplet. Participants also could indicate that they did not entertain any hypothesis. As in Kleinmann's (1993) study, participants were told that their responses would not be used to evaluate their performance.

At the end of the application training, participants were introduced to the dimensions of the integrity test and received back their ATIC questionnaire containing the hypotheses that they had entertained while filling out the integrity test. They were then asked to indicate for each of their hypotheses the integrity test dimension (also listed and explained on a separate sheet) to which it corresponded the most (cf. Kleinmann, 1993) or to indicate that a hypothesis did not correspond to any of the dimensions. However, they were not told what the correct dimension was.

In addition (and as a refinement of the procedure of Kleinmann, 1993), participants were also asked to rate the degree to which their hypothesis corresponded to this dimension (on a scale from 1 = *fits somewhat* to 4 = *fits completely*). Ratings of hypotheses corresponding to the correct dimensions were used as the measure of ATIC. If a dimension was not identified, a score of 0 was assigned. Thus, ATIC values could range between 0 (= no fit with the correct dimension) to 4 (= perfect fit with the correct dimension). For example, one participant wrote down "whether someone can keep a cool head" as his hypothesis when reading items belonging to the subscale Calmness. He later (correctly) indicated that his hypothesis corresponded best to the dimension Calmness and rated the correspondence as "4" (i.e., "fits completely"). When asked about his hypothesis when reading items belonging to the subscale Explanations for Illegitimate Behavior, he indicated that he had had "no hypothesis". Thus, he received 4 points for the Calmness triplet and 0 points for the Explanations for Illegitimate Behavior triplet. The mean of the correspondence ratings across all item triplets was used as the candidates' ATIC score.

Results

Descriptive information and correlations between the variables from this study are shown in Table 1. In line with our predictions, ATIC significantly correlated with participants' scores in the integrity test as a whole, $r = .23$, $p < .05$, and for the overt part of the integrity test, $r = .27$, $p < .01$. In contrast to this, the correlation between ATIC and scores in the personality-based part of the integrity test was not significant, $r = .05$, $p = .61$.

Table 1:
Descriptive Information, Correlations, and Cronbach's Alphas

	M	*SD*	1	2	3	4	5	6
1. Integrity score	3.74	.34	*.64*					
2. Integrity score – overt part only	3.72	.39	.94 [.91, .96]	*.91*				
3. Integrity score – personality-based part only	3.80	.38	.81 [.73, .87]	.57 [.42, .69]	*.93*			
4. ATIC (whole integrity test)	1.85	.90	.23 [.03, .41]	.26 [.06, .44]	.12 [-.08, .31]	*.63*		
5. ATIC (integrity test – overt part only)	1.31	.96	.25 [.05, .43]	.27 [.08, .45]	.14 [-.06, .33]	.92 [.88, .95]	*.41*	
6. ATIC (integrity test – personality-based part only)	2.94	1.20	.13 [-.07, .32]	.15 [-.05, .34]	.05 [-.15, .25]	.80 [.71, .86]	.50 [.33, .64]	*.59*

Note. ATIC = score for participants' ability to identify criteria. N = 92. The 95% confidence intervals are shown in brackets. Cronbach's Alphas appear in italics in the diagonal.

A potential limitation raised by a reviewer was that participants themselves rated the fit of their hypotheses to the test dimensions, which might have introduced a bias in our ATIC measure. We therefore additionally asked a recent graduate (with a Master's degree in work and organizational psychology) to rate all hypotheses. After being introduced to the dimensions, she read all hypotheses that participants had written down and rated them with the same rating scheme as the participants.[2] These externally rated ATIC scores showed the same correlational pattern with the integrity scores as the self-rated ATIC scores: $r = .24$, $p < .05$, for the integrity test as a whole, $r = .21$, $p < .05$, for the overt part only, and $r = .13$, $p = .21$, for the personality-based part.

Discussion

The aim of this study was to test whether ATIC correlates with integrity scores, and the results show that this was indeed the case. The finding that ATIC correlates with scores in an integrity test is important because it adds a new finding to the still small knowledge base on ATIC. Until now, we had known that ATIC correlates with AC performance (Kleinmann, 1993; Preckel & Schüpbach, 2005) and interview performance (Melchers et al., 2004), but the relationship between ATIC and integrity test scores had not been studied. Taken together with the other findings in the literature, this study shows that ATIC consistently correlates with scores of different kinds of personnel selection procedures.

Surprisingly, however, ATIC was only a significant predictor for the overt part of the integrity test and not for the personality-based part. This finding is at odds with evidence that overt scales from integrity tests are more fakable than personality-based scales (Alliger & Dwight, 2000). When we explored possible reasons for the differing correlations in our study, it turned out that candidates had considerably higher ATIC scores for the personality-based part than for the overt part (cf. Table 1), meaning that it was much easier for participants to identify the correct dimension for the former than for the latter, $t_{paired}(91) = 14.00$, $p < .001$, Cohen's $d = 1.46$. A possible reason for this difference could be that the labels and definitions of the overt scales are more abstract than the personality-based scales (e.g., Explanations for Illegitimate Behavior is more abstract than Calmness), and it is therefore probably more difficult to formulate hypotheses that really fit to the overt dimensions. Nevertheless, the personality-based part was also not completely transparent to participants (and also had greater variance than ATIC scores for the overt part), meaning that the differences between the correlational results cannot be attributed to ceiling or range restriction effects. Therefore, future research – possibly employing a conventional personality inventory instead of an integrity test – would appear to be necessary to establish whether the null relationship between ATIC and (personality) test scores is replicated.

It is a strength of this study that ATIC was measured similarly to previous studies (Kleinmann, 1993; Melchers et al., 2004) because this makes the results comparable with regard to the impact of participants' ATIC on their performance in various different selection procedures. At the same time, however, it is also a limitation, because ATIC in the integrity test is measured rather indirectly. Also, in order to simplify the candidates' task of finding a

[2] A subsample was coded by another rater. Eighty-five per cent of the hypotheses were assigned to the identical test dimension by both coders and the correlation between the two fit ratings was $r = .94$.

common label for their hypotheses for the item triplets used to measure ATIC, we had only selected positively worded items. Perhaps, however, these limitations could be preventable, because it might not be necessary to ask participants for their hypotheses regarding what integrity items measure. Instead, one could also present them with the test items and ask them a question like: "Do you think applicants should generally agree with this item to present themselves in a positive way?" Developing such a new ATIC measure seems a worthwhile task for future research.

Conducting this research made us aware that some integrity subscales have items that are similar to scales in personality tests but are scored in the reversed way. Thus, someone with a high ATIC should be able to identify that the modesty item in the integrity test should be strongly endorsed, whereas a modesty item in a personality test should not be endorsed. Such an applicant must take the context in which such items are embedded into account to understand that he should present himself differently depending on the kind of test. Do applicants really use such sophisticated strategies like taking into consideration the context of items? We do not yet know the answer to this question, but consider it a worthwhile research topic.

We should acknowledge that our arguments imply a causal relationship between ATIC and integrity scores (i.e., ATIC predicting integrity scores) even though our concurrent study design merely shows a correlation between ATIC and integrity scores. Experimental studies that manipulate the transparency of the dimensions in an integrity test are needed to show that there is a causal flow from ATIC to integrity scores (similar to the above-described studies by Kleinmann et al., 1996, and König et al., 2005).

Future research should also investigate the role of ATIC for the predictive validity of integrity tests. Kleinmann and colleagues (e.g., Kleinmann, 1993; Melchers et al., 2004) as well as Preckel and Schüpbach (2005) argue that individuals with high ATIC should not only perform better in an nontransparent AC but also in their daily job because ATIC can be considered an important social skill. Those people who can correctly interpret cues in their environment can also influence social interactions (e.g., with supervisors, colleagues, or customers) in a way that is satisfactory for all. Thus, ATIC in nontransparent personnel selection procedures should predict later performance on the job.

ATIC might also have an indirect effect on predictive validity, which should be explored in future studies. If applicants have a high ATIC and have, for example, figured out what the evaluation criteria are in an interview, they have three options: (a) They could give an answer that intentionally does not fit the criteria. If an applicant chose to do this, he would be deliberately lessening his chances of getting the job because non-fitting answers will lead to lower scores. We consider this option as extremely rare and therefore not worthy of further attention. (b) A more likely option is that people use their knowledge of the evaluation criteria for answering in a truthful way. They figure out that a biographical question is about leadership and come up with an appropriate leadership situation from their past. This should allow for a precise assessment of the construct in question (leadership in this example). In the general framework of validity (e.g., Binning & Barrett, 1989), it is believed that predictive validity should improve if constructs are better assessed. Thus, ATIC might have a positive indirect effect on predictive validity. (c) Some applicants might use their knowledge of what the evaluation criteria are (knowledge they have due to high ATIC) to exaggerate their positive attributes or for purposes of lying (i.e., faking). Lying in particular might introduce variance that is negatively related to job performance. Thus, ATIC could also have a negative indirect effect on predictive validity. Whether or not the negative effect of ATIC on

predictive validity is large depends on how many applicants actually exaggerate their attributes or tell lies. US-American research has so far shown that many applicants are fairly honest (Donovan, Dwight, & Hurtz, 2002), and this may be even more the case in Europe (Hafsteinsson, 2006). However, it is up to future research to show whether ATIC has a positive or a negative indirect effect on the predictive validity of nontransparent personnel selection procedures.

Taken together, this study reveals that ATIC – the ability to identify evaluation criteria – correlates with integrity scores. More generally, it supports the idea that ATIC is an important variable for personnel selection research.

References

Alliger, G. M., & Dwight, S. A. (2000). A meta-analytic investigation of the susceptibility of integrity tests to faking and coaching. Educational and Psychological Measurement, 60, 59-72.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. Personnel Psychology, 44, 1-26.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. Journal of Applied Psychology, 74, 478-494.

Connelly, B. S., Lilienfeld, S. O., & Schmeelk, K. M. (2006). Integrity tests and morality: Associations with ego development, moral reasoning, and psychopathic personality. International Journal of Selection and Assessment, 14, 82-86.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2002). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. Human Performance, 16, 81-106.

Franke, G. H. (1997). "The whole is more than the sum of its parts": The effects of grouping and randomizing items on the reliability and validity of questionnaires. European Journal of Psychological Assessment, 13, 67-74.

Fortmann, K., Leslie, C., & Cunningham, M. (2002). Cross-cultural comparisons of the Reid Integrity Scale in Latin America and South Africa. International Journal of Selection and Assessment, 10, 98-108.

Hafsteinsson, L. G. (2006). The prevalence of faking among Icelandic job applicants. Paper presented at the 21st annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. Journal of Applied Psychology, 78, 988-993.

Kleinmann, M., Kuptsch, C., & Köller, O. (1996). Transparency: A necessary requirement for the construct validity of assessment centres. Applied Psychology: An International Review, 45, 67-84.

König, C. J., Klehe, U.-C., Richter, G. M., Kleinmann, M., & Melchers, K. G. (2005). Transparency in structured interviews: Consequences for construct- and criterion-related validity. Paper presented at the 20th annual meeting of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

Marcus, B. (2006). Inventar berufsbezogener Einstellungen und Selbsteinschätzungen [Job-related attitudes and self-evaluations inventory]. Test manual. Göttingen: Hogrefe.

Marcus, B., Funke, U., & Schuler, H. (1997). Integrity Tests als spezielle Gruppe eignungsdiagnostischer Verfahren: Literaturüberblick und metaanalytische Befunde zur Konstruktvalidität [Integrity tests as a specific group of instruments in personnel selection: A literature review and meta-analytic findings on construct validity]. Zeitschrift für Arbeits- und Organisationspsychologie, 41, 2-17.

Marcus, B., Schuler, H., Quell, P., & Hümpfner, G. (2002). Measuring counterproductivity: Development and initial validation of a German self-report questionnaire. International Journal of Selection and Assessment, 10, 18-35.

McFarland, L. A., Ryan, A. M., & Ellis, A. (2002). Item placement on a personality measure: Effects on faking behavior and test measurement properties. Journal of Personality Assessment, 78, 348-369.

Melchers, K. G., Kleinmann, M., Richter, G. M., König, C. J., & Klehe, U.-C. (2004). Messen Einstellungsinterviews das, was sie messen sollen? Zur Bedeutung der Bewerberkognitionen über bewertetes Verhalten [Do selection interviews measure what they intend to measure? The impact of applicants' cognitions about evaluated behavior]. Zeitschrift für Personalpsychologie, 3, 159-169.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. Journal of Applied Psychology, 78, 679-703.

Preckel, D., & Schüpbach, H. (2005). Zusammenhänge zwischen rezeptiver Selbstdarstellungskompetenz und Leistung im Assessment-Center [Correlations between receptive self-presentation competence and performance in an assessment center]. Zeitschrift für Personalpsychologie, 4, 151-158.

Sackett, P. R., Burris, L. R., & Callahan, C. (1989). Integrity testing for personnel selection: An update. Personnel Psychology, 42, 491-529.

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.

Van Iddekinge, C. H., Taylor, M. A., & Eidson, C. E., Jr. (2005). Broad versus narrow facets of integrity: Predictive validity and subgroup differences. Human Performance, 18, 151-177.