Transparency in Structured Interviews: Consequences for Construct and Criterion-Related Validity

Ute-Christine Klehe
Universiteit van Amsterdam

Cornelius J. König
Universität Zürich

Gerald M. Richter
Philipps-Universität Marburg

Martin Kleinmann and Klaus G. Melchers
Universität Zürich

Although researchers agree on the use of structured interviews in personnel selection, past research has been undecided on whether these interviews need to be conducted nontransparently (i.e., without giving interviewees any indication of the evaluated criteria) or transparently (i.e., by revealing to interviewees the dimensions assessed in the interview). This article presents two independent studies examining the effects of interview transparency on interviewees' performance and on the interview's construct and criterion-related validity in the context of an application training program. Results from both Study 1 ($N = 123$) and Study 2 ($N = 269$) indicate an improvement in interviewees' performance under transparent interview conditions. Both studies further support the assumption that transparent interviews show satisfactory construct validity, whereas nontransparent interviews do not. Moreover, Study 2 showed no significant difference between the interview's criterion-related validity under transparent versus nontransparent conditions. Implications and directions for future research are discussed.

The prevalence of interviews for selecting job candidates (Dipboye, 1994) and the importance of structuring interviews so as to ensure reasonable psychometric properties (e.g., Conway, Jako, & Goodman, 1995; Huffcutt & Arthur, 1994) are beyond doubt. Most structured interviews are either past oriented, asking interviewees about their past experiences and behavior (e.g., patterned behavior description interviews; Janz, 1989), or future oriented, asking interviewees how they would respond to specific hypothetical scenarios (e.g., situational interviews; Latham, 1989).

A feature that has not received sufficient attention is the transparency with which interviews are conducted, that is, the degree to which applicants are told about the behavioral dimensions that an interview aims to assess. Some authors argue that interviewees should not learn the dimensions targeted by an interview (e.g., Latham & Saari, 1984; Latham & Sue-Chan, 1999; S. D. Maurer, Sue-Chan, & Latham, 1999). Some structured interviews, however, make the requirements of the interview questions clear (Huffcutt, Conway, Roth, & Klehe, 2004; Taylor & Small, 2002), either in the description of the situation ("Think about a time when you had to *motivate an employee* …" [italics added]) or in the actual question ("What would you do to *motivate the employee* to perform this task?" [italics added]; examples from Pulakos & Schmitt, 1995, p. 292). These diverging practices underline the necessity for research examining the consequences of interview transparency on interviewees' performance and on the construct and criterion-related validity of the structured interviews themselves. In this article, we argue why such effects are likely to exist and why they should be studied empirically.

INTERVIEW TRANSPARENCY

Transparency in structured interviews can be defined as the degree to which interviewees are informed about the particular requirements posed by the interview's questions. Thus, we regard transparency as a continuum ranging from highly nontransparent to highly transparent interviews. A nontransparent interview would be one during which interviewees are not given any indication whatsoever as to the dimensions or behavioral criteria by which their answers will be evaluated. Greater transparency can be induced if interviewees are informed in advance what kind of criteria the organization is looking for and which criteria will be assessed during the interview in general. Practitioners may provide participants with simple labels of these dimensions ("We're evaluating your teamwork, management skills, and leadership") or a short explanation as to what each of these labels stands for to ensure a correct understanding of the requirements by interviewees. Finally, practitioners may mention the required dimension directly prior to or as

part of each individual question, as carried out by Pulakos and Schmitt (1995) in the examples just presented.

There are different reasons why practitioners may conduct interviews transparently. Besides different traditions in personnel selection, it could be argued that transparent interviews are likely to facilitate the somewhat unnatural interaction with interviewees. If applicants know what is being assessed through a selection procedure, they can actively control the amount and nature of information they share, which may increase perceived procedural fairness (Gilliland, 1993; Stone & Stone, 1990). Transparency in selection interviews may also reduce interviewees' uncertainty and enable them to demonstrate their job-related knowledge, skills, and abilities (Arvey & Sackett, 1993; Schuler, 1993), thus reducing random error in the assessment of these job-related knowledge, skills, and abilities.

Transparency has to be distinguished from another intervention, namely coaching, defined as "any external intervention designed to improve scores" (Sackett, Burris, & Ryan, 1989, p. 147). Coaching differs from transparency in both its objectives and its implementation: It serves a different purpose (improving scores) than does transparency (fairness, measurement properties) and is often more general in focus rather than addressing only one particular set of interview questions. Although coaching may also involve the identification and explanation of dimensions usually targeted, it normally entails a prolonged training session during which interviewees learn about different types of interviews, general test-taking strategies on how to prepare and behave during the interview (keep eye contact, follow directions, time management, etc.), special strategies appropriate for certain types of interviews, and, if applicable, the logistics and rules of the particular interview they are to face in the near future. The main component of the interview coaching interventions discussed in the literature are interview role-plays for participants to practice and observe an actual interview including feedback on their own performance and a discussion of the exact scoring guidelines employed (e.g., Sackett et al., 1989). The coachings described by T. Maurer, Solamon, and Troxel (1998) and T. Maurer, Solamon, Andrews, and Troxel (2001), for example, offered a 2-hr session containing all of the aforementioned components open to all candidates up for promotion in the police and fire departments of a large city. Irrespective of the applicants' participation in the coaching session, however, the interview was also made transparent to all applicants by providing them with a "review of the interview knowledge, skill, and ability list, which was also distributed at the time of application" (T. Maurer et al., 1998, p. 130). Despite this transparency across interviewees, candidates participating in the coaching session still scored significantly better in the interview than those who had rejected the offer of coaching. This effect of coaching regardless of the interview's transparency supports the notion that coaching and transparency are not interchangeable constructs.

Consequences of Interview Transparency
on Interviewee Performance

When entering a selection interview, be it structured or unstructured, interviewees are likely to attempt to do well (e.g., Eder & Harris, 1999; Motowidlo, 1999). Consequently, knowledge of the dimensions targeted is likely to enhance interviewees' performance no matter whether this knowledge is the result of good decoding (cf. Kleinmann, 1993) or of explicit information about these dimensions. Nontransparent interviews force interviewees to attempt to discern the dimension targeted by each question, and their answers may therefore suit, neglect, or even oppose these dimensions. Given that interviewers, on the other hand, make their ratings based on preestablished scoring guides developed to reflect precisely these dimensions, interviewees who misjudge the purpose of a question will likely be given an inferior rating. A transparent interview, in contrast, leaves no room for such ambiguity. Interviewees are explicitly made aware of the targeted dimensions and face a strong situation with direct cues regarding the desired direction of their answer. Given that interviewees are likely to attempt to do well in the interview, they are more likely to shape each answer according to the dimension tackled in the interview question and should consequently, on average, receive higher ratings than interviewees facing a nontransparent interview.

   To our knowledge, no prior interview research has yet tested this assumption. For assessment centers (ACs), both Kleinmann, Kuptsch, and Köller (1996) and Smith-Jentsch (1996) reported that applicants performed better if they were informed about the targeted dimensions than if they were not informed (but see Kolk, Born, & van der Flier, 2003). Transparency should be more likely to influence results in interviews than in ACs, as interviewees only have to state the respective behavior. Although the high-fidelity simulation AC also demands proof of the applicants' ability to display the respective behavior, structured interviews only ask participants to report their past or their intended behavior without them actually having to act on it (Motowidlo, Dunnette, & Carter, 1990).

> H1: Interviewees who have been informed about the dimensions targeted in a structured interview (transparent condition) will receive better evaluations than interviewees who have not been informed about these targeted dimensions (nontransparent condition).

Consequences of Interview Transparency
on Construct Validity

Arguably, the measurement method of "interview" can be designed to reflect the characteristics of any specific position (Bobko, Roth, & Potosky, 1999; Campion, Palmer, & Campion, 1997). The number and type of constructs targeted across in-

terviews varies considerably (Huffcutt, Conway, Roth, & Stone, 2001). However, structured interviews suffer from a similar "validity paradox" as do ACs (Arthur, Woehr, & Maldegen, 2000; Klimoski & Brickner, 1987; Sackett & Dreher, 1982; Schneider & Schmitt, 1992). Although most structured interviews demonstrate content and criterion-related validity (e.g., Huffcutt et al., 2004; Taylor & Small, 2002), numerous studies have shown that such validity can hardly be attributed to the interviews' capacity to address those dimensions that the interviews had been developed to address, as the internal construct-related validity of interviews has been found to be notoriously low (Conway & Peneno, 1999; Huffcutt, Weekley, Wiesner, Degroot, & Jones, 2001; Schuler & Funke, 1989; Van Iddekinge, Raymark, Eidson, & Attenweiler, 2004): Interview questions developed to assess the same dimension often fail to load onto a common factor (resulting in weak dimension factors, i.e., in low convergent validity), and interview questions developed to assess completely different dimensions usually load onto a common factor if they belong to the same type of interview question (resulting in strong method factors, i.e., in low discriminant validity). Given the importance of construct validity for the general framework of validity (Binning & Barrett, 1989; Cascio, 1998) and for numerous administrative decisions regarding selection, placement, and training, it appears important and overdue to search for interventions that improve the low construct validity of interviews, ensuring that questions actually address the particular constructs that each of them had been developed to address.

Transparency may be a facet of standardization that improves an interview's measurement properties, such as its construct validity. During a nontransparent interview, interviewees may or may not identify which behavioral dimension a question targets (cf. Kleinmann, 1993). After misjudging the situation, their answers may not fit the intended dimension, irrespective of whether they actually possess the respective dimension-related abilities. Introducing random error into the assessment of applicant characteristics, this could decrease the interview's convergent validity: Questions that had been developed to assess the same dimension may be interpreted differently by interviewees and consequently may fail to correlate highly with one another or to represent a common factor.

In addition, during nontransparent interviews, the type of interview questions may overly impact on interviewees' reactions, independently of the questions' content, resulting in low discriminant validity: For example, focusing on their past versus imagining future scenarios, interviewees may fail to discern that different questions aim to assess different constructs within a particular type of interview.

Such difficulties should decrease during a transparent interview. If interviewees know that a question assesses, for example, leadership, they will be likely to try to show the leadership qualities that they possess, reducing the measurement error brought about by misjudging the situation. Indeed, they will be likely to react in the same way for other questions assessing leadership, leading to a more consistent and arguably more accurate assessment of their leadership abilities. Statistically,

this effect should be apparent in increased convergent validity, with interview questions increasingly loading onto those dimensions that they had been developed to assess.

Knowledge of the dimensions targeted may prevent interviewees from focusing solely on the different types of interview questions (e.g., whether an interview is directed toward the past or the future) and could allow them to differentiate their answers on the basis of the different dimensions the interview has been developed to assess, thus increasing the interview's discriminant validity.

H2: Transparency improves the internal construct validity of structured interviews.

Consequences of Interview Transparency on Criterion-Related Validity

Although transparency might foster the construct validity of structured interviews, there is even less agreement regarding criterion-related validity. Criterion-related validity might be of secondary concern if interviews primarily serve purposes of construct validity, such as placement decisions that match the requirements of a position with employees' strengths and weaknesses. For personnel selection, however, an interview's criterion-related validity is of primary concern.

The general framework of validity (Binning & Barrett, 1989; Cascio, 1998; Schleicher, Day, Mayes, & Riggio, 2002) suggests that interviews' criterion-related validity should improve with interview transparency. If interviewees' erroneous interpretations in nontransparent interviews distract them from answering in line with the required dimensions, this will introduce error into the measurement of those applicant characteristics that had been identified as relevant in the job analysis. Interview transparency, therefore, should increase the interviews' criterion-related validity by eliminating such erroneous interpretations and the resulting measurement error in dimension ratings.

At the same time, this argument assumes that interviewees' interpretations are random and unrelated to success in the job. Kleinmann (1993), however, showed that individuals differ reliably in their ability to identify criteria in nontransparent situations. He subsequently suggested that this ability to identify criteria in nontransparent situations might be as important during most performance settings under regular working conditions as it is during the personnel selection process. Conducting a selection procedure transparently would consequently suppress the assessment of an ability that might have contributed to the procedure's criterion-related validity.

Another argument assumes that when demands are not obvious to interviewees, an interview, and in particular a situational interview, will provide an opportunity to assess interviewees' motivation in the form of their intentions (e.g., Latham &

Saari, 1984; Latham & Sue-Chan, 1999; S. D. Maurer et al., 1999) or choices (e.g., Janz, 1989). Making interviewees explicitly aware of an interview's requirements, however, might tempt them to say whatever they were made to believe the interviewers wanted to hear—even though these answers might bear no relation to the behavior exhibited once interviewees actually face the situations described (Levashina & Campion, 2006).

Indirect empirical support for a hindering effect of transparency on the interviews' criterion-related validity again emerges in two studies on high-fidelity simulations. Kleinmann (1997) and Smith-Jentsch, Salas, and Brannick (2001) independently correlated performance on a predictor (an AC and a simulation exercise, respectively) with performance on a nontransparent criterion (another AC and self-evaluations, respectively) and found correlations to be higher when the predictor was nontransparent than when it was transparent.

Given these conflicting theoretical arguments, it appears more reasonable to address the effect of interview transparency on criterion-related validity with a research question rather than with a directed hypothesis.

RQ1: Will a transparent interview exhibit higher or lower criterion-related validity compared to a nontransparent interview?

We conducted two independent studies. Study 1 addressed H1 and H2, and Study 2 additionally tested RQ1.

STUDY 1

Methods

Sample.   This study was conducted as a job interview training program for university graduates who were currently applying for a job or would soon do so. The training was offered by the psychology department and the placement service of a German university together with a local branch of the German Federal Employment Office. To ensure that only participants motivated to do well took part in the training program, participants had to pay a small fee. The 123 participants (64 women, 59 men) had either just finished their master's degree (52 participants; 42.2%), were about to finish it (48 participants; 39.0%), or were studying for their undergraduate degree (22 participants; 19.7%). On average, participants were 27.8 years old ($SD = 5.37$) and had studied for 4.56 years ($SD = 2.15$). Fifty-one participants (41.5%) studied business administration, 28 participants (22.8%) various social sciences or arts, 22 participants (17.9%) natural sciences or engineering, and 8 participants (6.5%) law.

The setting of an application training program was chosen for three reasons: First, we needed a sample that was motivated to do well in the interviews, which excludes pure laboratory studies and concurrent organizational samples (Jennings, 1953). Second, a concurrent organizational sample was also deemed unfit for testing the hypotheses as it was necessary to choose a sample that, just like true applicants, had not yet gained extensive knowledge about the "organization" before entering the interviews. The use of a concurrent sample would have weakened the transparency manipulation because results for the nontransparent interview condition would have been confounded by interviewees' past experience with the respective job and organization (e.g., Depolo, Fraccaroli, & Sarchielli, 1994; Louis, 1980). A semicommercial training session for enhancing interview performance is likely to be the next best option after a sample of true applicants. Third, the use of a sample of true applicants would not have been ethical given our hypotheses regarding the influence of our manipulation on the interviewees' scores as well as the interviews' validity.

To enhance the realism of the situation and to give participants some indication as to what the "organization" might be looking for, participants had received a fictitious job advertisement about a management-trainee position at a nationwide car dealer and had been asked to prepare a written application for this position. We had chosen a management trainee position as it represented a realistic and attractive position for university graduates from diverse academic backgrounds.

Interviewers.   Most of the 25 interviewers were graduate students specializing in work and organizational psychology. All interviewers participated in half a day of interviewer training to become acquainted with the components and dimensions of the interviews. The training was designed to achieve a homogeneous understanding of the rating scales and anchors in order to reach a common frame-of-reference among interviewers. Interviewers also received information about typical rating errors. Interviewers conducted a complete trial interview with another interviewer acting as the interviewee.

Interviews.   Two subject-matter experts collected work-related incidents relevant for management trainees and deduced nine relevant behavioral dimensions (Flanagan, 1954). These dimensions were introduced to ten graduate work and organizational psychology students who had been interviewers in earlier studies during a workshop following the procedure outlined by Kleinmann et al. (1996). Students rated each dimension's ability to be rated during structured interviews as well as its conceptual independence from every other dimension. Although participants considered all nine dimensions to be able to be rated, they rated systematic planning, leadership, information management, and cooperation to be the dimensions most independent from one another. These applied mental and social skills

reflect four of the five constructs most frequently assessed in structured selection interviews (Huffcutt, Conway, et al., 2001).

Next, past- and future-oriented interview questions were formed from the original incidents (cf. Latham, 1989) or were taken from structured interviews developed and pretested for the banking (e.g., Schuler, 1988; Schuler & Moser, 1995) and engineering sectors (Deller & Kleinmann, 1993). During a final workshop, eight interview-experienced graduate work and organizational psychology students assigned each question to one of the four dimensions, excluding questions that could not be matched.

The final interview consisted of three components: a self-introduction, 12 past-oriented questions, and 12 future-oriented questions. The self-introduction stemmed from Schuler (Schuler, 1989a; Schuler & Funke, 1989), who argued that it could facilitate the rather uneasy interaction common in structured interviews (Campion et al., 1997; Latham & Finnegan, 1993) by giving interviewees the opportunity to reveal more about themselves. All interviewees were asked to "Take five minutes to introduce yourself. Please make special reference to your occupational history, your work-related strengths and weaknesses, and your aspirations." Interviewees' answers and behaviors were recorded and evaluated on each of the four dimensions on observation sheets similar to those used in ACs: Each dimension was described by positive and negative indicators from interviewees' answers (e.g., "indicates that he/she has happily and successfully worked in teams in the past") and behavior (e.g., "uses the time allotted without running over"). As is the case in traditional structured interviews (Campion et al., 1997), the panel of interviewers asked all interviewees the exact same questions, took notes, and scored responses according to a preestablished scoring scheme ranging from 1 (*unacceptable*) to 5 (*outstanding*). As frequently carried out in past-oriented interviews (e.g., Janz, 1989), interviewers asked for clarification or addition if an answer was unclear or incomplete. The self-introduction was based on a job analysis and had been pilot tested in the banking sector (Schuler, 1988). Testing the self-introduction as part of a "multimodal interview" during the selection of 500 bank clerk apprentices, Schuler and Moser (1995) found its criterion-related validities for supervisory performance assessments and assessments of potential 1 and 2 years later to be comparable to those of the more conventional structured interview components included in the multimodal interview.

The past-oriented questions, three for each of the assessed dimensions, asked interviewees to remember specific situations from their past and to describe their actions in these situations. For instance, a sample item addressing systematic planning was "You'll certainly remember your undergraduate exams. You had to revise two years of material and reproduce it in short sequences. How did you handle the load of material that you had to learn?" The scoring guide to this question suggested a score of 5 (*outstanding*) for an answer along the lines of "Made a schedule (when to study which topic, …). Developed goals and priorities for each subject.

Organized a suitable study group and researched the required learning materials in time. Did not experience serious time pressure"; a score of 3 (*acceptable*) for an answer along the lines of "Made a schedule, although I didn't always follow it or had to alter it. Experienced some time pressure (e.g., due to incomplete documents)"; and a score of 1 (*unacceptable*) for an answer along the lines of "Didn't make a schedule. Didn't set learning priorities. Experienced severe time pressure at the end of the revision period."

The future-oriented questions, again three for each dimension, confronted interviewees with hypothetical situations and asked them to describe what they would do in these situations. A sample item for leadership is "Imagine that you become the new head of department in a branch office. Yet, your new employees don't have much confidence in you and behave a little restrained. What would you do?" The scoring guide for this question suggested a score of 5 (*outstanding*) for an answer like "Seeks regular contact with employees (management by walking around), encourages them to approach him/her and talk about their work"; a score of 3 (*acceptable*) for an answer like "Is always willing to help them when they turn to him/her with questions"; and a score of 1 (*unacceptable*) for an answer like "Does nothing" or "Is convinced that with time, his/her technical competence will convince them."

Within each type of interview, items addressing the different constructs were presented in a randomized order that remained stable across all participants. In total, the self-introduction and the past and future-oriented questions took approximately 45 min to administer. As is common in interview studies (e.g., Latham & Skarlicki, 1995; Schuler & Funke, 1989), a panel of two interviewers asked each participant all interview components in direct succession. One interviewer read the questions and both interviewers recorded the interviewees' answers and rated them immediately.

After the interview, the interviewers discussed ratings on which they disagreed by 2 or more points on the 5-point scale. Most differences could be resolved quickly, although there was no requirement for interviewers to agree with one another. In the meantime, participants answered several demographic questions and a manipulation check regarding the transparency manipulation. Finally, they received individual feedback on their performance in the interview.

Transparency manipulation.   Both participants and interviewers had been assigned to either the nontransparent condition ($n = 64$) or the transparent condition ($n = 59$) based on their availability for the respective training. They were not aware of the other condition, the experimental manipulation, and the true purpose of the study (double-blind study).

Transparency was manipulated in the same way as it had been done in earlier studies (Kleinmann, 1997; Kleinmann et al., 1996; Kolk et al., 2003). Participants in the transparent condition received an introduction and a handout with defini-

tions of each dimension. The information on each dimension was relatively general and not particularly tailored to the interviews (e.g., "Leadership: takes on/aspires to leadership roles, takes charge, takes on role as group coordinator/speaker, argues for own point of view"). Meanwhile, participants in the nontransparent condition attended a short session about general application topics such as conducting job searches via the Internet and structuring application papers before entering the interview without information about the dimensions targeted by each question.

Results and Discussion

Preliminary analyses.   After the interviews, participants in the transparent condition answered six questions on a 5-point Likert scale ranging from 1 (*no*) to 5 (*yes*) (internal consistency = .71): "Were you aware of the dimensions assessed?"; "Did you try to present yourself in line with the dimensions assessed during the interview?"; and "Did you attempt to present yourself as a systematic planner/as an active leader/as capable of managing information/as cooperative?" Participants largely agreed with these items ($M = 3.86$, $SD = .58$), indicating that they had tried to present themselves in line with the dimensions.

Participants in the nontransparent condition answered two questions on a 5-point Likert scale ranging from 1 (*no*) to 5 (*yes*): "Did you wonder during the interview what the interviewers were trying to assess?" ($M = 3.75$, $SD = 1.27$) and "Did it appear obvious to you what the interviewers were trying to assess?" ($M = 3.13$, $SD = .89$). Participants agreed more with the first than with the second statement, $t(62) = 3.68$, $p < .01$, indicating that they felt relatively uninformed about the interviews' dimensions. We chose these differing manipulation checks in the two transparency conditions to reflect the fact that not only had participants been informed or not informed about the targeted dimensions but that the differing information had actually impacted on their cognitions and behavior during the interview.

Finally, we asked all participants how realistic they had perceived the interview situation to be: "Did you perceive the presented situations as realistic?" and "Did you try to react as you would also do on the job/… during a real job application?" answered on a 5-point Likert scale ranging from 1 (*no*) to 5 (*yes*). Participants indicated the interview to have been fairly realistic ($M = 3.69$, $SD = .90$), and their answers to have been very realistic ($M = 4.33$, $SD = .36$), with no differences emerging between the two experimental conditions on any of the three items, $t(121) = -.81$ to .65, *ns*. These answers suggest some generalizability of participants' behavior during the interview to true applicant samples.

Internal consistencies and interrater reliabilities are depicted on the left-hand side of Table 1. Except for nontransparent information management questions, internal consistencies were in the range to be expected given the interviews' structure and length (Conway et al., 1995). The low internal consistency of information

TABLE 1
Internal Consistencies and Interrater Reliability for Interviews in Study 1 and Study 2

| | Study 1 | | | | | | Study 2 | | | | | |
| | Internal Consistencies | | | Interrater Reliability | | | Internal Consistencies | | | Interrater Reliability | | |
	Total[a]	Nontransparent[b]	Transparent[c]	Total[a]	Nontransparent[b]	Transparent[c]	Total[d]	Nontransparent[e]	Transparent[f]	Total[d]	Nontransparent[e]	Transparent[f]
Interview by dimensions												
Systematic planning	.68	.62	.68	.86	.82	.87	.66	.61	.66	.82	.82	.82
Leadership	.68	.47	.78	.89	.83	.91	.69	.66	.70	.87	.86	.86
Information management	.56	.21	.71	.85	.81	.86	—	—	—	—	—	—
Cooperation by component	.57	.46	.65	.80	.83	.76	.59	.55	.55	.82	.79	.81
Self-introduction	.79	.78	.76	.87	.88	.84	.71	.74	.65	.83	.81	.82
Past-oriented questions	.63	.34	.76	.86	.81	.88	.70	.68	.70	.85	.84	.85
Future-oriented questions	.79	.69	.83	.90	.89	.87	.71	.68	.71	.85	.83	.85
Complete interview	.85	.70	.89	.91	.86	.91	.83	.80	.82	.89	.89	.88
Criterion by dimension												
Systematic planning	—	—	—	—	—	—	.62	.66	.60	.91/.71[g]	.88/.66[g]	.93/.74[g]
Leadership	—	—	—	—	—	—	.58	.62	.54	.94/.80[g]	.92/.79[g]	.94/.78[g]
Cooperation by exercise	—	—	—	—	—	—	.52	.66	.60	.92/.73[g]	.88/.72[g]	.93/.71[g]
Group Discussion 1	—	—	—	—	—	—	.89	.86	.90	.90	.86	.92
Group Discussion 2	—	—	—	—	—	—	.86	.87	.85	.87	.82	.90
In-basket	—	—	—	—	—	—	.60	.67	.53	.97	.96	.98
Complete criterion	—	—	—	—	—	—	.84	.84	.82	—	—	—

[a]$N = 123.$ [b]$n = 64.$ [c]$n = 59.$ [d]$N = 269.$ [e]$n = 110.$ [f]$n = 159.$ [g]Interrater reliabilities are divided because the number or observers per exercise differ: Numbers before the slash indicate reliabilities in the group discussions (four observers), numbers after the slash indicate reliabilities in the in-basket (two observers).

management in the nontransparent condition could not be attributed to any specific item but pertained across all seven measurement points of this dimension.

H1. H1 postulated that interviewees score better in structured interviews if they are informed about the targeted dimensions. Means and standard deviations of the studied variables are presented in the left columns (nontransparent condition) and upper rows (transparent condition) of Table 2. We tested H1 via a one-way multivariate analysis of variance (MANOVA) with overall ratings of systematic planning, leadership, information management, and cooperation as the dependent variables and transparency condition as the independent variable. Results revealed a significant effect of transparency condition on interviewee performance, Wilks's $\lambda = .88$, $F(4, 118) = 4.15$, $p < .01$, partial $\eta^2 = .12$, thus supporting H1. Additional analyses of variance (ANOVAs) on individual dimension ratings supported this finding for all dimensions with $F(1, 121) = 12.84$, $p < .01$, partial $\eta^2 = .10$ for planning; $F(1, 121) = 12.06$, $p < .01$, partial $\eta^2 = .09$ for leadership; $F(1, 121) = 10.13$, $p < .01$, partial $\eta^2 = .08$ for information management; and $F(1, 121) = 5.44$, $p = .02$, partial $\eta^2 = .04$ for cooperation.

H2. H2 postulated that structured interviews' internal construct validity would improve if interviewees learned about the targeted dimensions. As in earlier analyses of interview construct validity (e.g., Schuler, 1989b), the different dimensions were treated as different traits, the different interview formats as different methods in the following analyses. The nontransparent condition showed a mean monotrait–heteromethod (MTHM; convergent) correlation of .20 and a mean heterotrait–monomethod (HTMM; discriminant) correlation of .30, as compared to a mean heterotrait–heteromethod (HTHM) correlation of .10. This indicates both poor convergent and poor discriminant construct validity. In the transparent condition, these correlations were .39 (convergent), .46 (discriminant), and .32, respectively, indicating a considerable increase in convergent validity yet also a decrease in discriminant validity. MTHM correlations surpassed HTMM correlations 69 times in the nontransparent condition and 74 times in the transparent condition, suggesting a slight increase in construct validity in the transparency condition (Campbell & Fiske, 1959) and thus tentative support for H2.

STUDY 2

Study 1 supported our notion that making structured interviews transparent to interviewees would (a) raise interviewees' performance in the interview and (b) result in an increased construct validity. At the same time, above multitrait–multimethod (MTMM) approach has several difficulties, including in particular a lack of criteria for quantifying the degree of construct validity. A more elegant ap-

TABLE 2
Study 1: Means, Standard Deviations, and Correlations Among Interview Dimensions

	M	SD	Self-Introduction				Past-Oriented Interview				Future-Oriented Interview			
			SYST	LEAD	INFO	COOP	SYST	LEAD	INFO	COOP	SYST	LEAD	INFO	COOP
M			3.82	3.50	3.66	3.83	4.04	3.83	3.84	3.68	3.80	3.83	3.70	3.76
SD			1.09	1.26	1.00	.96	.60	.96	.72	.81	.72	.80	.66	.78
Self-introduction														
SYST	3.46	.93		.55**	.56**	.50**	.17	.35**	.18	.32*	.29**	.13	.44**	.14
LEAD	2.74	1.07	.49**		.50**	.14	.27*	.69*	.29*	.27*	.39**	.38**	.41**	.29*
INFO	3.11	.93	.51**	.44**		.42**	.19	.42**	.38**	.33*	.22	.28*	.35**	.37**
COOP	3.34	.88	.45**	.49**	.45**		.12	.07	.07	.20	.22	-.05	.23	.27*
Past-oriented interview														
SYST	3.73	.65	.40**	.11	.27*	.24		.31*	.56**	.42**	.50**	.47**	.41**	.43**
LEAD	3.29	.80	.22	.40**	.09	.29*	.08		.46**	.33*	.43**	.52**	.43**	.29*
INFO	3.81	.50	.18	.10	.28*	.15	.20	.25*		.38**	.44**	.56**	.51**	.45**
COOP	3.70	.75	-.01	-.03	.17	.14	.04	-.16	.12		.48**	.35**	.51**	.34**
Future-oriented interview														
SYST	3.34	.65	.38**	-.02	.25*	.11	.37*	-.01	.23	.16		.52**	.50**	.50**
LEAD	3.26	.82	.25*	-.05	.22	.15	-.11	.10	.16	.23	.32**		.50**	.48**
INFO	3.36	.51	.19	.07	.06	.09	-.01	.07	.11	.12	.29*	.24		.43**
COOP	3.53	.77	-.06	-.21	-.02	-.05	.01	-.15	.07	.24	.35**	.39**	.40**	

Note. Descriptives and correlations in the top right-hand corner represent the transparent condition ($n = 59$). Descriptives and correlations in the bottom left-hand corner represent the nontransparent condition ($n = 64$).

SYST = systematic planning; LEAD = leadership; INFO = information management; COOP = cooperation.

$*p < .05.$ $**p < .01$ (two-tailed).

proach for testing H2 would be the use of confirmatory factor analyses (CFAs), which require a larger sample size than available for Study 1, however. Findings from Study 1 also suggested that a second study should strive for a higher reliability per dimension, a stronger transparency manipulation, and a larger sample size. Study 2 therefore served to replicate our findings and to test RQ1.

Methods

The procedure was the same as in Study 1, with the following exceptions: First, we enhanced the experimental manipulation of transparency versus nontransparency. Second, RQ1 required the additional collection of criterion data. Finally, to establish high and comparable reliabilities among the assessed dimensions (Marsh, 1989), information management, which had suffered from severe reliability problems in the nontransparent condition of Study 1, was removed from the interviews in favor of revising the interviews and using four instead of three questions for each of the remaining three dimensions in both the past- and the future-oriented interviews.

Sample. The 269 participants (138 women, 131 men) were an average of 28.46 years old ($SD = 5.68$) and had studied for 4.70 years ($SD = 1.78$). Seventy-six participants (28.3%) studied business, 19 participants (7.1%) law, 42 participants (24.9%) social sciences, and 35 participants (13%) natural sciences or engineering. More than half of the sample (55.4%) reported having work experience, with a range from 0.5 to 22 years ($M = 4.88$ years, $SD = 5.21$). Consequently, 74.7% had experience in applying for actual jobs.

The interviews. The interviews in Study 2 were developed to assess the same mental and social skills systematic planning, leadership, and cooperation that had already been targeted in Study 1. The selection and refinement of items followed the same procedure as in Study 1, albeit with different interview-experienced master's-level work and organizational psychology students serving as subject matter experts. The final interview again consisted of the self-introduction, followed by 12 past- and 12 future-oriented questions. Within the past- and the future-oriented types of interview questions, items addressing the different constructs were presented in a randomized order that remained stable across all participants.

Performance criterion. The criterion for this study had to satisfy at least five conditions: First, it should reflect the same job-related content and constructs that were targeted in the interview to ensure that (a) criterion-related validity was not influenced by a lack of fit between interview and criterion and (b) the criterion represented actual behavior. Second, the criterion should be of comparable nontrans-

parency across participants. Although the dimensions observed in the criterion should not be specifically revealed in order to reflect the fact that organizational employees often disagree as to the behaviors relevant for performance and promotion (Beehr & Taber, 1993), we wanted to ensure a comparable level of experience and information about the criterion across participants. Third, the assessment should be practicable and prevent participant attrition. Fourth, different observers should rate participants' performance in the interview and the criterion to prevent common observer biases. Finally, given the aforementioned effects of interview transparency on interviewee performance, we felt it inadequate to conduct this study in a field setting. Consequently, as in the study by Kleinmann (1997), our study used several nontransparent high-fidelity simulations (Motowidlo et al., 1990) as the criterion. Although it is no assessment of actual performance on a job, such a proxy criterion assesses performance in high-fidelity simulations of situations deemed representative and critical for the respective position.

The proxy criterion consisted of three high-fidelity simulations deemed particularly suitable for assessing the required dimensions by 3 personnel selection experts and 12 master's-level work and organizational psychology students who had been observers and/or interviewers in earlier training studies. These simulations were two leaderless group discussions, one with and one without assigned roles, and an in-basket exercise in which participants had to organize and schedule numerous events. The inclusion of the in-basket task required workshop participants to match each of the in-basket's original evaluation criteria to one of the three targeted dimensions (e.g., "Participant recognizes collision between appointments A and B" became part of systematic planning), enabling observers to evaluate participants' in-basket performance on all three dimensions. Each group discussion was evaluated by four observers, each in-basket by two observers. Again, performance was scored between 1 (*unacceptable*) and 5 (*outstanding*) on each of the three dimensions.

Procedure.   Once again, master's-level work and organizational psychology students served as interviewers and observers. They had been prepared through an intensive one-day training course similar to the one in Study 1 with the addition of learning how to assess participants' performance in the proxy criterion.

The day prior to the actual training, participants attended a 1-hr meeting in which they received the training-schedule, general information about the exercises, and some information about the job in the form of a fictitious job advertisement for a management trainee position. To ensure comparable nontransparency of the proxy criterion across conditions, it was assessed prior to the interviews. Finally, participants answered a number of demographic and manipulation check questions. Observers were rotated between the proxy criterion and the interview to ensure that their knowledge of participants' performance in the proxy criterion

could not influence the ratings participants received in the interview. Interviewers had no opportunity to discuss their impressions with criterion observers.

Experimental manipulation.   Both interviewers and participants were assigned to either the nontransparent ($n = 110$) or the transparent condition ($n = 159$) based on their availability for the training and were unaware of any experimental manipulation, the other condition, and the study's true purpose (double-blind study). Participants in the nontransparent condition entered the interview without any information regarding the dimensions. Participants in the transparent condition underwent a stronger manipulation than participants in Study 1 to mirror the degree of transparency of interviews that directly state the targeted dimension in each question (e.g., Pulakos & Schmitt, 1995). Besides receiving the same introduction and handout with definitions as in Study 1, interviewees were informed before each question about the specific dimension targeted (e.g., "The following question addresses leadership").

Results and Discussion

Preliminary analyses.   As in Study 1, following the interview, participants in the transparent condition answered three questions, namely, "Did you think that the instruction prior to each question made you understand the dimension assessed?" "Did you think that the instruction prior to each question made you understand the scoring standards employed?" and "Did you try to present yourself in line with the assessed dimensions during the interview?" Answered on a 4-point Likert scale ranging from 1 (*no*) to 4 (*yes*), the internal consistency of the scale was .57. Participants indicated that they had largely understood and acted upon the targeted dimensions ($M = 3.30$, $SD = .48$).

Participants in the nontransparent condition received a list of eight potential dimensions and were asked to indicate which dimension had been assessed by each interview question. On average, participants correctly identified systematic planning in 54% ($SD = .28$), cooperation in 37% ($SD = .23$), and leadership in 42% ($SD = .24$) of the respective questions, indicating that the dimension assessed by each question was reasonable but not strikingly obvious or transparent to participants.

Finally, three questions addressed all participants' perceived realism of the interview situation: "Did you perceive the training as a realistic reflection of an application situation?", "Were you able to assume the role of an applicant?" and "Did you try to react as you would also do during a real job application?" answered on a 4-point Likert scale ranging from 1 (*no*) to 4 (*yes*) (internal consistency = .65). On average, participants confirmed the realism of the interview situation and of their reactions ($M = 3.22$, $SD = .52$), with no differences emerging between experimental conditions, $t(266) = .41$, *ns*. These answers indicate some generalizability of our results to applicant samples. Internal consistencies and interrater agreement

for the different constructs assessed, for the types of interview questions employed, and for the proxy criterion in Study 2 are depicted on the right-hand side of Table 1.

H1.   H1 postulated that interviewees score better in structured interviews if they are informed about the targeted dimensions. Means and standard deviations of the studied variables are presented in the left columns (nontransparent condition) and upper rows (transparent condition) of Table 3. A one-way MANOVA with transparency condition as the independent variable and systematic planning, leadership, and cooperation as the dependent variables was used to test this hypothesis. In line with H1, the MANOVA revealed a significant effect of transparency condition on interviewee performance, Wilks's $\lambda = .89$, $F(3, 265) = 11.15$, $p < .01$, partial $\eta^2 = .11$, thus supporting H1. Additional ANOVAs on separate dimension ratings supported this finding for all dimensions with $F(1, 267) = 18.71$, $p < .01$, partial $\eta^2 = .07$ for systematic planning; $F(1, 267) = 20.04$, $p < .01$, partial $\eta^2 = .07$ for leadership; and $F(1, 267) = 28.44$, $p < .01$, partial $\eta^2 = .10$ for cooperation.

H2.   H2 postulated that the structured interviews' internal construct validity would improve with interview transparency. The nontransparent condition showed a mean MTHM (convergent) correlation of .24 and a mean HTMM (discriminant) correlation of .41, as compared to a mean HTHM correlation of .15, again indicating poor convergent and poor discriminant validities. In the transparent condition, these correlations were .29 (convergent), .39 (discriminant), and .15, respectively, indicating a slight improvement in both convergent and discriminant validity. Similarly, the number of times that a MTHM correlation surpassed an HTMM correlation (Campbell & Fiske, 1959) was 11 in the nontransparent and 17 in the transparent condition, respectively, indicating a first tentative support for H2.

For a concise test of the interview's construct validity, we additionally conducted CFAs. CFAs also allow comparisons between multiple latent models that could account for interview ratings. Like Van Iddekinge et al. (2004), we employed two types of CFA models to test the interviews' construct validity. The first CFA model assumes correlated traits and correlated methods (CTCM) and can be characterized as follows: (a) Nine observed variables measure (b) three different traits (i.e., dimensions), represented by three latent trait factors, with three different methods (i.e., types of interview questions), represented by three latent method factors. (c) The trait factors can correlate with each other and the method factors can correlate with each other, but correlations between trait and method factors are zero. (d) Each observed variable loads onto only one trait factor and onto only one method factor. (e) The error terms ("uniquenesses") of observed variables are uncorrelated among themselves but can be freely estimated. Thus, a CTCM measures the influence of the dimensions assessed (traits) and the influence of types of interview questions used (methods).

TABLE 3
Study 2: Means, Standard Deviations, and Correlations Among Interview Dimensions and Criteria

	M	SD	Self-Introduction			Past-Oriented Interview			Future-Oriented Interview			Criterion		
			SYST	LEAD	COOP	SYST	LEAD	COOP	SYST	LEAD	COOP	SYST	LEAD	COOP
M			3.78	3.37	3.92	3.60	3.40	3.49	3.60	3.45	3.81	3.23	2.88	3.26
SD			0.83	1.00	0.86	0.61	0.78	0.60	0.63	0.58	0.56	.59	.63	.59
Self-introduction														
SYST	3.49	.81		.45**	.38**	.20*	.19*	.09	.15	.13	.10	.19*	.19*	.11
LEAD	2.73	1.05	.51**		.32**	.20*	.43**	.18*	.20*	.24**	.09	.23**	.29**	.15
COOP	3.50	.97	.47**	.49**		.24*	.20*	.21**	.27**	.17*	.21**	.19*	.15	.16*
Past-oriented questions														
SYST	3.36	.68	.02	.11	.24*		.30**	.38**	.43**	.38*	.35**	.29**	.21**	.28**
LEAD	3.19	.85	.30**	.40**	.45**	.28**		.40**	.41**	.32**	.22**	.31**	.28**	.16*
COOP	3.27	.59	.10	.24*	.31**	.45**	.36**		.27**	.32**	.38**	.29**	.19*	.17*
Future-oriented questions														
SYST	3.34	.69	.22*	.17	.24*	.31**	.38**	.33**		.45**	.44**	.30**	.29**	.25**
LEAD	3.31	.57	.28**	.12	.28**	.10	.31**	.37**	.43**		.38**	.30**	.28**	.26**
COOP	3.45	.65	.04	-.15	.15	.22*	.09	.32**	.27**	.42**		.25**	.22*	.20*
Criterion														
SYST	3.14	.59	.26**	.25**	.20*	.20*	.24*	.09	.28**	.22*	.16		.80**	.66**
LEAD	2.28	.63	.24*	.20*	.27**	.22*	.29**	.25**	.35**	.28**	.21*	.76**		.62**
COOP	3.06	.59	.15	.18	.26**	.27**	.28**	.36**	.30**	.30**	.23*	.69**	.66**	

Note. Descriptives and correlations in the top right-hand corner represent the transparent condition ($n = 159$). Descriptives and correlations in the bottom left-hand corner represent the nontransparent condition ($n = 110$). SYST = systematic planning; LEAD = leadership; INFO = information management; COOP = cooperation.

*$p < .05$. **$p < .01$ (two-tailed).

Unfortunately, the CTCM approach frequently suffers from statistical problems (see Tomás, Hontangas, & Oliver, 2000), the most crucial being the production of poorly defined solutions: Solutions do not reach convergence, parameter estimates are outside their permissible range (e.g., negative error variances), or models are underidentified. In response to these problems, Marsh (1989) proposed the correlated-trait correlated-uniqueness (CTCU) approach as an alternative CFA approach for analyzing construct validity. Unlike the CTCM approach, the CTCU approach does not require separate method factors but rather infers method effects from correlated uniquenesses within methods (i.e., correlations among error terms of performance measures within the same method). Although the CTCU approach rarely produces ill-defined solutions, it has its own methodological and theoretical disadvantages (see Lance, Noble, & Scullen, 2002; Tomás et al., 2000), which have caused researchers to suggest the use of both CTCM and CTCU approaches when testing for construct validity (Conway, 1996; Lance et al., 2002).

Following the procedure outlined by Byrne (1994), both approaches served to test three different models: Within the CTCM approach, the first model (Model 1a) contained only three methods (i.e., types of interview) and no trait (i.e., dimension) factors, thus abandoning the notion of dimensions. Model 2a included the same three methods (i.e., types of interview) factors and one general trait (i.e., dimension), representing the idea that interviewers could not distinguish between dimensions. Model 3a is the traditional CTCM model with three dimensions and three method factors.

Within the CTCU approach, Model 1b allowed only correlated errors (correlated uniquenesses) within methods (i.e., types of interviews). Model 2b added one general trait (i.e., dimension) factor, and Model 3b included all three dimensions instead of only one. As outlined by Byrne (1994), a significant difference in fit between Models 3 and 1 proves the interviews' convergent validity, and a significant difference in fit between Models 3 and 2 demonstrates the interviews' discriminant validity.

We tested the proposed models with AMOS 5 (Arbuckle, 2003), using the covariance matrix as the input for the program. Chi-square statistics, the root mean square error of approximation, the comparative fit index, the Tucker-Lewis index, the incremental fit index, the goodness-of-fit index, and the adjusted goodness-of-fit index were used to assess the fit of the models. Following Hu and Bentler (1999), sample size and model complexity suggest that goodness-of-fit indexes should surpass .95 and root mean square error of approximation should remain below .06 to indicate a relatively good fit.

Table 4 shows the results of the CFAs for the transparent condition, and Table 5 shows the results those for the nontransparent condition. The CTCM approach yielded nonadmissible solutions for Model 3a under both conditions because of poorly defined solutions (e.g., negative error variances). Such estimation problems are typical for CTCM models (Tomás et al., 2000) and render them uninterpretable.

TABLE 4
Study 2: Summary of GFIs for Models—Transparent Condition

Model	df	χ^2	p	χ^2/df	RMSEA	CFI	TLI	IFI	GFI	AGFI	Admissible?
Correlated trait correlated methods models											
Model 1a: 3 correlated methods	24	41.55	.01	1.73	.07	.94	.91	.94	.95	.90	Yes
Model 2a: 1 dimension, 3 correlated methods	15	20.34	.21	1.27	.04	.99	.97	.99	.97	.93	No[a,b]
Model 3a: 3 correlated dimensions, 3 correlated methods	12	10.88	.54	.91	.00	1.00	.98	1.00	.96	.86	No[a,b]
Correlated trait correlated uniqueness models											
Model 1b: correlated errors within methods	27	128.76	.00	4.77	.15	.65	.54	.66	.85	.75	Yes
Model 2b: 1 dimension, correlated errors within methods	18	30.69	.03	1.71	.07	.96	.91	.96	.96	.90	Yes
Model 3b: 3 correlated dimensions, correlated errors within methods	15	18.14	.26	1.21	.04	.99	.97	.99	.98	.93	Yes

Note. $n = 159$. RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; IFI = incremental fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index.
[a]Model does not converge. [b]Model has negative error variance(s).

TABLE 5
Study 2: Summary of GFIs for Models—Nontransparent Condition

Model	df	χ^2	p	χ^2/df	RMSEA	CFI	TLI	IFI	GFI	AGFI	Admissible?
Correlated trait correlated methods models											
Model 1a: 3 correlated methods	24	55.47	.00	2.31	.11	.86	.79	.87	.89	.80	Yes
Model 2a: 1 dimension, 3 correlated methods	15	27.73	.02	1.85	.09	.94	.86	.95	.95	.84	Yes
Model 3a: 3 correlated dimensions, 3 correlated methods	12	17.38	.14	1.45	.06	.98	.93	.98	.97	.88	No[a]
Correlated trait correlated uniqueness models											
Model 1b: correlated errors within methods	27	100.48	.00	3.72	.16	.67	.56	.68	.84	.73	Yes
Model 2b: 1 dimension, correlated errors within methods	18	36.79	.01	2.04	.10	.92	.83	.92	.93	.83	Yes
Model 3b: 3 correlated dimensions, correlated errors within methods	15	34.90	.00	2.33	.11	.91	.79	.92	.94	.81	Yes

Note. $n = 110$. RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; IFI = incremental fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index.
[a]Theta-delta matrix (error covariance matrix) of the model is not positive definite.

However, all CTCU models converged. In the nontransparent condition, Model 3b, proposing three dimensions besides the correlated uniquenesses, yielded a significantly better fit than Model 1b, which only assumed correlated uniquenesses, $\Delta\chi^2(12) = 65.58$, $p < .01$, indicating that the interviews possessed convergent validity. However, a comparison between Model 3b and the more parsimonious Model 2b, which assumes one instead of three dimension factors, turned out to be not significant, $\Delta\chi^2(3) = 1.89$, $p > .10$. As Model 2b and 3b also yielded no acceptable fit (see Table 5), this indicates that the nontransparent interviews showed no proof of discriminant validity.

In the transparent condition, Model 3b yielded a better fit than both Model 1b, $\Delta\chi^2(12) = 110.62$, $p < .01$, and Model 2b, $\Delta\chi^2(3) = 12.55$, $p < .01$, indicating that the interviews possessed both convergent and discriminant validity. As Model 3b also yielded excellent fit indices (see Table 5), one can infer the interviews' good internal construct validity under the transparent condition. H2 was therefore supported in that the same interview that lacked construct (particularly discriminant) validity when administered nontransparently did exhibit such validity when it was administered transparently. The parameter estimates for the respectively best fitting models in both conditions are depicted in Table 6.

Post Hoc Analyses

An alternative possible explanation for the aforementioned results might be that interviewees in the transparency condition may be better able to tell interviewers what they want to hear.[1] Although this explanation might well account for the increased level of performance in the transparency condition (H1), it does not account for the increase in construct validity (H2) in both studies, as there is little reason to assume that interviewees are better at telling interviewers what they want to hear so for some dimensions than for others—unless one assumes that each interviewee has a better grasp of some dimensions than of others, which again confers with the proposed assumption that transparency reduces measurement error and thus improves the relevance of interviewees' answers.

To further test empirically, that changed ratings were not due to social desirability, we collected self-reported social desirability from participants in the transparency condition. One hundred thirty-six participants filled out the Marlowe-Crowne Social Desirability Scale (Marlowe & Crowne, 1961), the most commonly used assessment of social desirability bias. Sixty-eight of these participants also filled out the Balanced Inventory of Desirable Responding (Paulhus, 1984), which addresses participants' conscious impression management and unconscious self-deception. None of the correlations between self-reported scale

---

[1] We thank an anonymous reviewer for raising this point.

TABLE 6
Study 2: Parameter Estimates for Best Fitting Models for the Nontransparent Condition
(CTCU Model With 1 Dimension and 3 Exercises) and for the Transparent Condition (CTCU Model With 3 Dimensions and 3 Exercises)

| | Nontransparent Condition | | | | | | | Transparent Condition | | | | | | | | |
| | Dimension Parameter | Latent Error Term | | | r Between Error Terms Within Methods | | | Dimension Parameter Loadings | | | Latent Error Term | | | r Between Error Terms Within Methods | | |
Variable	Loading	R^2	Variance	SE	SYST	LEAD	COOP	SYST	LEAD	COOP	R^2	Variance	SE	SYST	LEAD	COOP
Self-introduction																
SYST	.28*	.08	.60	.08				.28**			.08	.63	.07			
LEAD	.42**	.17	.91	.13	.45**				.47**		.22	.77	.10	.40**		
COOP	.50**	.25	.70	.11	.40**	.36**				.40**	.16	.62	.08	.33**	.26**	
Past-oriented questions																
SYST	.46*	.21	.36	.07				.72**			.51	.18	.05			
LEAD	.85**	.72	.20	.15	−.25				.78**		.61	.23	.09	−.43†		
COOP	.70**	.48	.18	.06	.20	−.61				.63**	.40	.22	.05	.01	.15	
Future-oriented questions																
SYST	.47**	.22	.37	.06				.59**			.35	.25	.04			
LEAD	.41**	.17	.26	.04	.30*				.49**		.24	.26	.04	.26*		
COOP	.22*	.05	.39	.05	.19	.37**				.55**	.30	.21	.03	.27*	.26**	
Variance (SE)	.10 (.05)							.19 (.06	.08 (.03)	.09 (.03)						
SYST																
LEAD								.87**								
COOP								.80**	.65**							

Note. $n_{(\text{nontransparent})} = 110$; $n_{(\text{transparent})} = 159$. CTCU = correlated-trait correlated-uniqueness; SYST = systematic planning, LEAD = leadership, COOP = cooperation.
*$p < .05$. **$p < .01$ (two-tailed). †This relatively large but nonsignificant ($p = .08$) effect is primarily due to the relatively large standard error of the respective covariance ($SE = .050$, as compared to .042 and .035 for the remaining two covariances in that set).

scores and performance in the transparent interviews reached even marginal significance, correlations being .05 (*ns*) for the Marlowe-Crown scale and .16 (*ns*) and .15 (*ns*) for the Social Desirability and Impression Management subscales of the Balanced Inventory of Desirable Responding, respectively, indicating that respondents' performance in the transparent interview was not a mere function of social desirable responding.

RQ1   RQ1 asked whether interviewees' performance in a nontransparent criterion would correlate higher or lower with their performance in a nontransparent or in a transparent interview. The lower rows (nontransparent condition) and right columns (transparent condition) of Table 3 present the correlations between the constructs assessed in the interviews and in the proxy criterion. The mean correlation between interview ratings and performance in the proxy criterion was .24 in the nontransparent and .22 in the transparent condition, respectively. We tested RQ1 by multiple regression of overall performance in the proxy criterion. After controlling for transparency condition ($\beta = .00$, *ns*) and performance in the overall interview ($\beta = .45$, $p < .01$), an answer for RQ1 is provided by the interaction term between transparency condition and interview performance. No such interaction emerged ($\Delta R^2 = .00$, *ns*). The same pattern of results was found for predicting systematic planning ($\Delta R^2 = .00$, *ns*), cooperation ($\Delta R^2 = .00$, *ns*), and leadership ($\Delta R^2 = .00$, *ns*) with the respective dimensions assessed in the interviews. Thus, transparency neither increased nor decreased the interviews' criterion-related validity.

GENERAL DISCUSSION

In this article we examined the effects of conducting structured interviews transparently on interviewees' performance and on the construct and criterion-related validity of structured interviews. Both studies suggest that interviewees perform better and that the interviews show improved internal construct validity when the interview was administered transparently. Finally, we found no effect of interview transparency on the interview's criterion-related validity.

Given the lack of research on the influence of interview transparency on interviewees' performance and the mixed findings reported in the related literature on ACs (Kleinmann, 1997; Kleinmann et al., 1996; Kolk et al., 2003; Smith-Jentsch, 1996), our study's finding of increased performance in the transparent condition highlights the difference between high- and low-fidelity simulations. Although interviews are usually more practicable and cost-efficient than ACs and are widely accepted among hiring organizations, they represent only low-fidelity simulations in which actual behavior needs to be inferred from interviewees' answers without interviewees proving that they are truly capable of enacting the required behavior (Motowidlo et al., 1990). Thus, knowledge of the requirements posed by a given

question is quite likely to influence interviewees' responses, a concern that could only partially be supported in the case of ACs.

The finding of improved performance in the transparent condition also highlights the importance of ensuring that any specific selection interview is equally transparent or nontransparent across interviewees. Clearly, this is not a difficult requirement in the case of transparent interviews. However, performance rankings may be significantly distorted as soon as some interviewees learn about the targeted dimensions, whereas others do not. This effect is likely to occur when organizations, especially prominent organizations approached by many applicants, maintain the same interview for an extended time. Some new applicants may learn about the questions and the scoring employed via former applicants or service agencies specializing in providing such information, thus unduly improving their performance, whereas other applicants may know little about the interview requirements. This also indicates that feedback about interviewees' performance during selection interviews may be detrimental to the interviews' validity for future applicants. Especially among prominent organizations, former interviewees may share the feedback they received with later applicants, and interviews may thus become transparent to some applicants while remaining nontransparent to others.

The importance of construct validity for many administrative decisions warrants further investigation of factors that may enhance it. To the authors' knowledge, transparency represents the first moderator found for interview construct validity: When administered transparently, the interviews showed construct validity, even though they failed to do so when administered nontransparently. A likely reason for this is a reduction in measurement error due to interviewees' incorrect interpretations of questions. Transparency may therefore help organizations to acquire a more accurate picture of the interviewees' relative strengths and weaknesses on the targeted dimensions. This makes transparent interviews a useful and relatively cheap tool for placement decisions or needs assessments regarding training interventions. They might also provide future or current employees with a reference as to what specific requirements the organization expects, what the respective behavior of each requirement looks like, and finally in which areas employees need to adapt to meet the organization's requirements.

At the same time, we had expected interview transparency to affect the interviews' criterion-related validity. The general framework of validity (Binning & Barrett, 1989; Schleicher et al., 2002) suggested that transparency improved interviews' criterion-related validity due to a reduction of error. Conversely, Kleinmann's (1993) idea of a stable ability to identify the requirements or social situations suggested that transparency would suppress an ability that otherwise contributed to the procedure's criterion-related validity. In terms of the relationships found in Study 2, neither effect emerged.

There could be different reasons for this: First, both conceptions may be stronger theoretically than practically so that actual effects may be weak. Second, it is

equally likely that both positive and negative effects of interview transparency on criterion-related validity do exist but cancel each other out. Although the increase in the interviews' construct validity might lead to an increase in criterion-related validity, transparent interviews may also suppress interviewees' ability to understand the workings of a situation and to respond accordingly. An assessment of this ability (Kleinmann, 1993) under both nontransparent and, for purposes of comparison, transparent conditions and its effect on the interviews' criterion-related validity, may shed more light onto this possible mechanism.

The discussion of the interviews' criterion-related validity leads us to the main limitations of the study. First, the ethical and methodological reasons just outlined prevented us from using applicants for an actual job. This appears to be less problematic for the interviews, as many participants were currently looking for jobs and were using the training program as a chance to prepare for actual interviews. Also, the posttraining questionnaires indicated that participants perceived the training program as realistically reflecting an actual application situation, eliciting realistic participant reactions. The setting of an application training program might have been more problematic in relation to the assessment of the proxy criterion in the form of three high-fidelity simulations judged as representative for the role of management trainees. The proxy criterion was assessed directly before the interviews, which may have enhanced the chance of finding increased relationships between the interviews and the proxy criterion due to shared error (e.g., because of participants having a good or a bad day). This setting may also reduce the external validity of the results obtained for RQ1. Finally, it is very possible that a nontransparent criterion does not reflect the practice of organizations that are highly structured and outspoken about their performance requirements.

In addition, we might have used more direct manipulation checks, even though this would have been more of a problem if none of the proposed effects had emerged between conditions. The results for H1, however, suggest that participants in the transparent condition had gained some advantage through the manipulation compared to participants in the nontransparent condition.

Finally, it should be noted that the interview was assessed in the same order (self-introduction, past-oriented questions, future-oriented questions) across all interviewees, and it is difficult to tell how this might have affected results. It could be argued that the self-introduction, which allows for a somewhat longer elaboration of answers than either the past- or the future-oriented questions, might influence results primarily of the past-oriented questions following directly afterward, so that differences in results on the past-oriented questions are partially due to the influence of the self-introduction. Similarly, Latham and Skarlicki (1995) administered the future-oriented questions prior to the past-oriented questions to ensure that knowledge or partial transparency gained in the past-oriented questions could not influence results on the future-oriented questions. If this had been a serious concern in our study, participants' performance on the future-oriented questions

would have been equally high in the nontransparent and transparent conditions, which was not the case in either of the two studies. Consequently, even though counterbalancing the types of interview questions would have been desirable, we do not expect this to have greatly altered results.

In sum, the results of these studies reveal that there could be some merit in making the targeted dimensions clear to interviewees. Such procedures are likely to be beneficial in terms of fairness towards applicants by raising the situational comparability among them. Future research should investigate the effect of transparent administration on interviewee reactions. It is also possible that transparency allows organizations to provide interviewees with feedback and to use the same interviews for a longer period. However, before a clear recommendation to practitioners to start administering their structured interviews transparently instead of nontransparently can be made, further research is needed, especially on the criterion-related validity of transparent versus nontransparent interview questions for predicting performance in various field settings (social, administrative, etc.).

ACKNOWLEDGMENTS

REFERENCES

Arbuckle, J. L. (2003). *AMOS* (5th ed.). Chicago: Smallwaters Corporation.

Arthur, W., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical reexamination of the assessment center construct-related validity paradox. *Journal of Management, 26*, 813–835.

Arvey, R. D., & Sackett, P. R. (1993). Fairness in selection: Current developments and perspectives. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection* (pp. 171–202). San Francisco: Jossey-Bass.

Beehr, T. A., & Taber, T. D. (1993). Perceived intra-organizational mobility: Reliable versus exceptional performance as means to getting ahead. *Journal of Organizational Behavior, 14*, 579–594.

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.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52*, 561–589.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/WINDOWS. Basic concepts, applications, and programming*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection inter-view. *Personnel Psychology, 50*, 655–702.

Cascio, W. F. (1998). *Applied psychology in human resource management* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Conway, J. M. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management, 22*, 139–162.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consis-tency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565–579.

Conway, J. M., & Peneno, G. M. (1999). Comparing structured interview question types: Construct va-lidity and applicant reactions. *Journal of Business and Psychology, 13*, 485–506.

Deller, J., & Kleinmann, M. (1993). Das situative Interview [The situational interview]. In A. Gebert & W. Hacker (Eds.), *Arbeits- und Organisationspsychologie 1991 in Dresden* (pp. 336–343). Bonn, Germany: Deutscher Psychologen Verlag.

Depolo, M., Fraccaroli, F., & Sarchielli, G. (1994). Le décalage entre attentes et réalité dans le processus de socialisation au travail [Mismatches between expectations and reality on the work so-cialization process]. *Travail Humain, 57*, 131–143.

Dipboye, R. L. (1997). Structured selection interviews: Why do they work? Why are they underutilized? In N. Anderson & P. Herriot (Eds.), *International handbook of selection and assess-ment* (pp. 455–473). New York: Wiley.

Eder, R. W., & Harris, M. M. (1999). Employment interview research. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 1–27). Thousand Oaks, CA: Sage.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327–358.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspec-tive. *Academy of Management Review, 18*, 694–734.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Con-ventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for en-try-level jobs. *Journal of Applied Psychology, 79*, 184–190.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U.-C. (2004). Evaluation and comparison of the sit-uational and behavior description interview formats. *International Journal of Selection and Assess-ment, 12*, 262–273.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic as-sessment of psychological constructs measured in employment interviews. *Journal of Applied Psy-chology, 86,* 897–913.

Huffcutt, A. I., Weekley, J. A., Wiesner, W. H., Degroot, T. G., & Jones, C. (2001). Comparison of situa-tional and behavior description interview questions for higher-level positions. *Personnel Psychology, 54*, 619–644.

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 158–168). Thousand Oaks, CA: Sage.

Jennings, E. E. (1953). The motivation factor in testing supervisors. *Journal of Applied Psychology, 37*, 168–169.

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

Kleinmann, M. (1997). Transparenz der Anforderungsdimensionen: Ein Moderator der Konstrukt- und Kriteriumsvalidität des Assessment-Centers [Transparency of the required dimensions: A moderator of assessment centers' construct and criterion validity]. *Zeitschrift für Arbeits- und Organisations-psychologie, 41*, 171–181.

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

Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40*, 243–260.

Kolk, N. J., Born, M. P., & van der Flier, H. (2003). The transparent assessment centre: The effects of revealing dimensions to candidates. *Applied Psychology: An International Review, 52*, 648–668.

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods, 7*, 228–244.

Latham, G. P. (1989). The reliability, validity, and practicality of the situational interview. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 169–182). Thousand Oaks, CA: Sage.

Latham, G. P., & Finnegan, B. J. (1993). Perceived practicality of unstructured, patterned, and situational interviews. In H. Schuler (Ed.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 41–55). Hillsdale, NJ: Lawrence Erlbaum Associates.

Latham, G. P., & Saari, L. M. (1984). Do people do what they say? Further studies on the situational interview. *Journal of Applied Psychology, 69*, 569–573.

Latham, G. P., & Skarlicki, D. P. (1995). Criterion-related validity of the situational and patterned behavior description interviews with organizational citizenship behavior. *Human Performance, 8*, 67–80.

Latham, G. P., & Sue-Chan, C. (1999). A meta-analysis of the situational interview: An enumerative review of reasons for its validity. *Canadian Psychology, 40*, 56–67.

Levashina, J., & Campion, M. A. (2006). A model of faking likelihood in the employment interview. *International Journal of Selection and Assessment*, *14*, 299–316.

Louis, M. R. (1980). Surprise and sense making: What newcomers experience in entering unfamiliar organizational settings. *Administrative Science Quarterly, 25*, 226–251.

Marlowe, D., & Crowne, D. P. (1961). Social desirability and response to perceived situational demands. *Journal of Consulting Psychology, 25*, 109–115.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13*, 335–361.

Maurer, S. D., Sue-Chan, C., & Latham, G. P. (1999). The situational interview. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 159–177). Thousand Oaks, CA: Sage.

Maurer, T., Solaman, J., Andrews, K. D., & Troxel, D. (2001). Interview coaching, preparation strategies, and response strategies in relation to performance in situational employment interviews. And extension of Maurer, Solamon, and Troxel (1998). *Journal of Applied Psychology, 86*, 709–717.

Maurer, T., Solaman, J., & Troxel, D. (1998). Relationships of coaching with performance in situational employment interviews. *Journal of Applied Psychology, 83*, 128–136.

Motowidlo, S. J. (1999). Asking about past behavior versus hypothetical behavior. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 179–190). Thousand Oaks, CA: Sage.

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

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598–609.

Pulakos, E. D., & Schmitt, N. (1995). Experience-based and situational interview questions: Studies of validity. *Personnel Psychology, 48*, 289–308.

Sackett, P. R., Burris, L. R., & Ryan, A. M. (1989). Coaching and practice effects in personnel selection. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 145–183). New York: Wiley.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67*, 401–410.

Schleicher, D. J., Day, D. V., Mayes, B., T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735–746.

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology, 77*, 32–41.

Schuler, H. (1988). *Handbuch: Standardisierungshilfen für Einstellungsgespräche. Ausbildungsberufe Bankkaufmann/ Sparkassenkaufmann* [Handbook: Standardizing selection interviews. Financial services clerks]. Stuttgart, Germany: Deutscher Sparkassenverlag.

Schuler, H. (1989). Construct validity of a multimodal employment interview. In B. J. Fallon, H. P. Pfister, & J. Brebner (Eds.), *Advances in industrial organizational psychology* (pp. 343–354). North-Holland: Elsevier.

Schuler, H. (1993). Social validity of selection situations: A concept and some empirical results. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 11–26). Hillsdale, NJ: Erlbaum.

Schuler, H., & Funke, U. (1989). The interview as a multimodal procedure. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 183–192). Thousand Oaks, CA: Sage.

Schuler, H., & Moser, K. (1995). Die Validität des Multimodalen Interviews [Validity of the multimodal interview]. *Zeitschrift für Arbeits- und Organisationspsychologie, 39*, 2–12.

Smith-Jentsch, K. A. (1996). *Should rating dimensions in situational exercises be made trnsparent for participants? Empirical tests of the impact on convergent and predictive validity.* Paper presented at the 11th annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Smith-Jentsch, K. A., Salas, E., & Brannick, M. T. (2001). To transfer or not to transfer? Investigating the combined effects of trainee characteristics, team leader support, and team climate. *Journal of Applied Psychology, 86*, 279–292.

Stone, E. F., & Stone, D. L. (1990). Privacy in organizations: Theoretical issues, research findings, and protection mechanisms. *Research in Personnel and Human Resource Management, 8*, 349–411.

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology, 75*, 277–294.

Tomás, J. M., Hontangas, P. M., & Oliver, A. (2000). Linear confirmatory factor models to evaluate multitrait-multimethod matrices: The effects of number of indicators and correlation among methods. *Multivariate Behavioral Research, 35*, 469–499.

Van Iddekinge, C. H., Raymark, P. H., Eidson, C. E., & Attenweiler, W. J. (2004). What do structured selection interviews really measure? The construct validity of behavior description interviews. *Human Performance, 17*, 71–93.