Situational Judgment Tests in High-Stakes Settings: Issues and Strategies With Generating Alternate Forms

Filip Lievens  
Ghent University

Paul R. Sackett  
University of Minnesota, Twin Cities Campus

This study used principles underlying item generation theory to posit competing perspectives about which features of situational judgment tests might enhance or impede consistent measurement across repeat test administrations. This led to 3 alternate-form development approaches (random assignment, incident isomorphism, and item isomorphism). The effects of these approaches on alternate-form consistency, mean score changes, and criterion-related validity were examined in a high-stakes context (N = 3,361). Generally, results revealed that even small changes in the context of the situations presented resulted in significantly lower alternate-form consistency. Conversely, placing more constraints on the alternate-form development process proved beneficial. The contributions, implications, and limitations of these results for the development of situational judgment tests and high-stakes testing are discussed.

Keywords: situational judgment tests, alternate forms, item generation, high stakes testing

An important aspect of many high-stakes testing programs is the need to develop alternate forms of the test in question. Whereas some high-stakes tests are designed for one-time use (e.g., a city administers a police officer selection test once every several years with the expectation that a completely new examination will be developed per administration), in many cases there is a need for periodic administration of comparable tests (e.g., educational admissions tests or occupational licensure and certification tests) at intervals that range widely, from monthly in some settings to annually in others. The need for alternate tests is motivated by concerns that reuse of the same test will result in a decline in the validity of the test due to a variety of mechanisms, ranging from test security breaches such that the initial test items become known to concerns that individuals who retest after an initial failure are unfairly advantaged by prior exposure to items (Sackett, Burris, & Ryan, 1989).

There exist well-developed technologies for addressing this problem, and conceptual and operational issues in the development of parallel tests have long been of interest to the measurement field (Nunnally & Bernstein, 1994). Central features of the classic version of these technologies include the development of item pools, the pretesting of items to gather item statistics, and the development of test construction approaches that impose various item parameter and content constraints (e.g., specifying the use of items matched on difficulty and discrimination parameters). In domains in which the features that determine item characteristics (e.g., item difficulty) are well known, recent developments in item generation theory (Irvine & Kyllonen, 2002) permit the development of items with known characteristics.

However, there are settings in which these approaches are not feasible. One key determinant is the feasibility of pretesting items to obtain the item statistics needed. In many cases, there are mechanisms for such pretesting, such as the embedding of new items in operational tests for research purposes, or the availability of appropriate populations of individuals to participate in testing sessions solely for purposes of obtaining item statistics. In other cases, there are constraints that prevent pretesting. One such constraint is test security. In some settings in which an employer is developing job knowledge examinations for promotional purposes, pretesting items on, for instance, an incumbent population is judged too great a threat to security. Another constraint may be legal. The current study involves a government-administered admissions examination for which including nonscored items for research purposes is prohibited. Thus, it is not uncommon to face the challenge of having to develop tests intended to be parallel to existing tests without the availability of item trialing. A second impediment to the classic psychometric approach is the scenario in which the predictor (a) is multidimensional and (b) reflects a construct domain that is not fully understood (Clause, Mullins, Nee, Pulakos, & Schmitt, 1998). Examples of such predictors are situational judgment tests (SJTs) or assessment centers. With these predictors, individual items and exercises are commonly designed to sample key job domain aspects, rather than to reflect a clearly understood construct.

We find it useful to cross these two features. Developing alternate forms is most straightforward when one can pretest items reflecting a unidimensional well-understood construct (see Nunnally & Bernstein, 1994) or use item generation theory to “clone” items, such that item characteristics are known without pretesting. Conversely, it is most challenging when one cannot pretest items...
and when the test is a multidimensional one that represents constructs not fully understood.

This study deals with this last scenario. Specifically, we focused on one multidimensional predictor (SJT) that has received increased interest in both the employment (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001) and educational domains (Lievens, Buyse, & Sackett, 2005a; Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004). We examined different approaches to developing alternate SJT forms in an operational context when item pretesting is not possible. At a conceptual level, we borrowed from principles underlying item generation theory by posing competing perspectives about which features of SJTs might enhance or impede consistent measurement across test administrations. In the next sections, we present item generation theory and apply it to examine the effectiveness of three approaches to generating alternate SJT items. We also formulate hypotheses about the effects of these approaches on key variables in high-stakes selection settings.

Study Background

**Item Generation Theory**

Item generation theory has emerged as a useful framework for producing comparable forms of tests of unidimensional constructs (e.g., verbal reasoning, numerical reasoning) without extensive item trialing (Bejar, 1986; Collis, Tapsfield, Irvine, Dann, & Wright, 1995; Embretson, 1998; Irvine & Kyløllon, 2002; Kyløllon, & Christal, 1990). The hallmark of item generation theory is that one a priori determines the factors that contribute to item difficulty (Kyløllon, 2002). Along these lines, the radicals–incidents approach is often used. Radicals refer to structural item features that determine item difficulty (Irvine, Dann, & Anderson, 1990). Radicals are also known as controlling factors (Dennis, Handley, Bradon, Evans, & Newstead, 2002). Conversely, incidents refer to changes in surface characteristics of items that do not determine item difficulty. Incidents are also called noncontrolling or nuisance factors (Dennis et al., 2002). For instance, in a visual tracking task, radicals might refer to the number of targets to be tracked, the speed of the moving targets, and the routes they take, whereas incidents might refer to the shape or the color of the targets. Similarly, in a verbal reasoning test, radicals might consist of the use of negations and class membership, whereas incidents might entail the use of different fonts or nouns.

Once the radicals and incidents (and their levels) are determined on the basis of expert judgments or cognitive theory, they are varied to produce (either automatically or manually) lists of item variants. Some of these item variants are isomorphs because they have the same radical but different incidents. Thus, they look superficially different, even though they are psychometrically equivalent. Other items are true variants because they differ in terms of both radicals and incidents (Bejar, 2002). A recent review of empirical studies using item generation theory found general support for its effectiveness (Irvine, 2002). Item difficulty for various test items could be well predicted from radicals. In addition, different forms constructed through item generation principles had a high degree of parallelism and were equally related to relevant criterion variables (e.g., training performance). Finally, the different approaches of generating items did not produce a change in the construct being measured.

In sum, at a theoretical level, item generation theory forces test developers to think about the factors that contribute to task difficulty. Accordingly, item generation theory might serve as a meta-item writing approach wherein the focus is placed on item specifications instead of on the items themselves, leading to various item design rules (Kyløllon, 2002). Practically, the generation of comparable forms might provide a shield against test compromise and guarantee test security in high-stakes test situations wherein item trialing is not possible.

**Three Perspectives About Radicals and Incidents in SJTs**

As noted in the preceding section, the essence of item generation theory is that one considers the factors that contribute to item difficulty and the ones that do not. Although this might be relatively straightforward for unidimensional tests due to the presence of theory and research, it is more challenging for the construction of alternate forms of a multidimensional test such as an SJT. In this respect, Kyløllon (2002) warned:

> How is that we know which factors are radical and which incidental? One approach, and it seems the most common one, is to designate certain item factors incidental a priori. . . . The problem is that these a priori incidentals could turn out to be important determinants of item difficulty. . . . This suggests that the radical-incidental distinction could at least be partially empirically determined. (p. 261)

Clearly, SJT items differ along a multitude of dimensions (McDaniel & Nguyen, 2001). So far, there does not exist any theory about how people answer SJTs or about what makes some SJT items more or less difficult than others. In addition, the constructs underlying SJTs are equivocal (see Schmitt & Chan, 2006, for a review). On the one hand, factor analytic research has typically revealed that various factors that are difficult to interpret explain about the same portion of variance. On the other hand, SJT items have been found to be correlated with a wide array of constructs such as general cognitive ability, personality, work experience, and job knowledge (McDaniel & Nguyen, 2001). Hence, the general consensus seems to be that SJTs are better conceptualized as methods that can be designed to measure a broad variety of constructs (depending on the domain/criterion of interest).

Absent a taxonomy or theory of SJTs, we decided to posit different perspectives about what might serve as radicals or incidents for SJTs. The first perspective proposes that the universe of SJT items varies along various features but that the major determining feature is the domain to be targeted. As shown in Table 1 (upper part), the domain is considered to be the only radical, whereas all other SJT features are assumed to provide content, context, and linguistic variation in the items. The underlying rationale is that all critical incidents related to a specific domain are merely variations on the same theme and will all provide information on the standing of a person on a given criterion construct of interest (e.g., interpersonal domain or customer service performance domain). Thus, an SJT that measures interpersonal dimensions will be highly correlated with another SJT that captures the same interpersonal domain. However, it will not be correlated with an SJT that measures goal setting.

To create alternate SJT forms, this first perspective can be operationalized using the following strategy. A large enough pool
of SJT items is developed with the only requirement that they tap the same domain. Next, these SJT items are randomly assigned to alternate forms. Hence, the remainder labels this strategy the random assignment strategy.

The second perspective posits that the defining feature of SJT items is not only the domain to be targeted but also the critical incidents that are gathered to develop the SJT items. As shown in Table 1, it is assumed that both the domain and the incidents are the radicals, whereas all other SJT features (e.g., the specific context wherein the critical incident and responses are embedded, linguistic and grammar changes to item stems and options) are assumed to be incidentals.

To create alternate SJT forms, this perspective implies that pairs of items that are designed to reflect the same critical incident (e.g., a physician dealing with a patient who refuses medication) are developed, with one of each pair then assigned to alternate forms. Thus, the critical incidents built into the items are held constant across alternate SJT forms. However, the concrete context wherein these incidents are embedded (the item stems) and the ways of responding to them (the item options) differ. In this study’s SJT, context refers among others to the background, characteristics, symptoms, and disease of a patient. As shown in Appendix A, the same critical incident about refusal to take medication might be operationalized in a context wherein a female patient who is an active runner refuses to take specific medication and in a context wherein a male patient refuses to take medication because people told him it was no good.

The third perspective posits that even relatively minor changes in the situations presented in multidimensional predictors such as SJTs might result in lack of consistency across alternate forms. As can be seen in Table 1, it is assumed that not only the domain and the critical incidents included, but also the context wherein they are embedded and their ways of responding to them, are radicals.

According to this perspective, incidentals in SJT items refer only to the linguistic and grammar changes to item stems and item options to be used.

To create alternate SJT forms, pairs of items that are designed to reflect the same domain, the same critical incident, and the same context of the item stem and responses are developed. In other words, the content and context similarity across pairs of items assigned to alternate forms is maximized. Only changes in terms of wording and grammar are allowed across the alternate items. This approach can be labeled as item isomorphic. In the only study we are aware of that used this procedure (which was then labeled item cloning), Clause et al. (1998) reported alternate-form reliabilities greater than .70 for three parallel forms of a 33-item SJT used in a research context. Yet the effectiveness of this approach has not been put under scrutiny in an actual selection context. When used in an actual high-stakes context with highly motivated participants, a potential unexamined drawback of the use of item isomorphic procedures might be that the items across examinations are so similar that the test becomes prone to large mean score changes. Appendix B presents an example of the item isomorphic strategy.

### Hypotheses

As mentioned in the above quote by Kyllonen (2002), it is important to empirically verify the radical–incidental distinction for a given test when there is no strong a priori or theoretical basis. Accordingly, it might become clear how similar test items have to be to yield consistent measurement across alternate forms. In this study, empirical verification about the controlling SJT features could be obtained by testing the influence of the various strategies on key dependent variables in high-stakes testing programs such as alternate-form consistency, robustness to mean score changes, and criterion-related validity.

As shown in Table 1, the three perspectives result in alternate SJT forms that vary in similarity. At one extreme, SJTs are considered to provide consistent measurement across alternate forms as long as the domain is similar. At the other extreme, SJTs are considered to provide consistent measurement across alternate forms as long as the domain, the incidents, and the context of incidents and item options are similar. We hypothesized that the degree of conceptual similarity of the SJT forms would be positively related to alternate-form consistency. Thus, we expected that less restrictive approaches of constructing alternate SJT forms...
such as random assignment might reduce the consistency of SJT scores across alternate SJT forms as the incidents, item stems, and response options were not held constant. Conversely, adding more constraints on alternate SJT development (e.g., incident isomorphic and item isomorphic approaches) might increase the alternate-form consistency of SJT scores. This led to the following hypotheses:

Hypothesis 1a: The correlation among two alternate SJT forms will be higher for the item isomorphic approach than for the incident isomorphic approach.

Hypothesis 1b: The correlation among two alternate SJT forms will be higher for the incident isomorphic approach than for the random assignment approach.

A second challenge in high-stakes selection is to minimize retest effects. Retest effects can be defined as mean score changes on a test after prior exposure to an identical test or to an alternate form of this test under standardized conditions. Using this definition, retest effects encompass both practice and coaching effects (Kulik, Kulik, & Bangert, 1984; Lievens, Buyse, & Sackett, 2005b; Messick & Jungeblut, 1981; Sackett et al., 1989). Prior research in the cognitive ability domain has shown that retest effects depend on the similarity between the original test and the retest. For instance, in the educational field Kulik et al. conducted a meta-analysis of retest effects related to cognitive ability tests. The average effect size was .42 for identical tests and .23 for parallel tests. Thus, the amount of prior exposure to the test seems to be a central determinant of retest effects.

We hypothesized that the degree of conceptual similarity of the SJT forms would be negatively related to the retest effects obtained. In the item isomorphic approach, candidates are presented with virtually the same items, some linguistic and grammar changes notwithstanding. In a high-stakes selection context wherein candidates are motivated to improve on their scores, the item isomorphic approach might lead to large retest effects. Conversely, in the random assignment approach, the items deal with different problem situations because the critical incidents built into the item stems are not kept constant. Hence, the random assignment model might be effective in minimizing retest effects. The incident isomorphic approach can be positioned somewhere in the middle of these two extremes in terms of producing retest effects. This led to the following hypotheses:

Hypothesis 2a: Retest effects will be larger for the item isomorphic approach than for the incident isomorphic approach.

Hypothesis 2b: Retest effects will be larger for the incident isomorphic approach than for the random assignment approach.

Finally, we considered the effects of the various approaches for constructing alternate SJT forms on criterion-related validity (i.e., whether a given person’s initial test or retest is more valid). Along these lines, Lievens et al. (2005b) distinguished between four possible mechanisms behind retest effects: (a) measurement error; (b) genuine improvement of the candidate’s standing on the characteristic; (c) criterion-relevant change in the observed score that reduces or eliminates a test deficit (e.g., unfamiliarity with the test) between the observed score and an unchanging true score; and (d) criterion-irrelevant (i.e., artificial) change on the characteristic of interest, resulting from the learning of tricks, recall of repeated items, or improper access to test content. As it is difficult to distinguish between these various mechanisms in an actual high-stakes context, we did not posit hypotheses or examine the validity of the different approaches for explorative purposes.

Method

The study was situated in the context of admission to medical college. Specifically, data were collected during the admission exams for medical studies administered between 1998 and 2005 in Belgium. Apart from the SJT, the exam consists of a science knowledge test, a cognitive ability test, and a silent reading protocol (Lievens et al., 2005a). Each year the examination is administered twice, with many individuals who failed on the first administration retaking the examination at the second administration. Thus, two examination forms are created each year (with a 2-month time interval), and the intent is that the two forms be comparable. Over the years, the development of the SJT has changed. In the early years of the exam, the two annual forms of the SJT were developed according to the random assignment model. After auditing the exam at the end of 1999, a decision was made to use the incident isomorphic model (see Lievens et al., 2005b). Finally, after 2003, the incident isomorphic model was replaced by the item isomorphic model (see Clause et al., 1998).

As we conducted this study in a high-stakes context, the three alternate-form construction approaches for the SJT were logically confounded with the time period. Hence, it is important that no other changes (e.g., demographic composition of the candidate pool, procedural changes) occurred across the years. As noted below, the admission exam remained the same across years on these key factors. One exception is that the presentation format of the SJT was changed from video-based to written (Lievens & Sackett, 2006). This change did not coincide with a change in alternate-form development procedure. In fact, the written format was introduced after 2002. Hence, written SJT versions were developed on the basis of both the incident and item isomorphic approaches. In any case, we analyzed whether the change in presentation format affected our results.

Given that this study focused on the effects of repeat administrations of the SJT, only candidates who participated twice in the admission exam in a given year were considered. Between 1998 and 2005, there were 3,361 candidates who participated twice in the admission exam. The demographic make-up of these candidates across the three periods of interest to this study was as follows: 703 candidates (36% male, 64% female; average age = 18.4 years) in the first time period, 1,385 candidates (33% male, 67% female; average age = 18.6 years) in the second time period, and 1,273 candidates (31% male, 69% female; average age = 18.4 years) in the third time period. As can be seen, the demographic characteristics of these samples were very similar across the three time periods of interest to this study, even though there was a trend that more female students participated in recent years. The fact that the demographic characteristics of these samples were virtually the
same is not surprising, because both of these samples represented
the population of candidate medical students (i.e., all students who
participated in a specific year in the admissions exam).

**Procedure**

Each year, the admission exam lasted for a whole day and was
centrally administered in a large hall. The administration of the
exam was highly standardized, as it was guided by a minute-by-
minute script. On average, the passing rate of the admission exam
was about 30%. One week after the exam, candidates obtained
feedback on their specific test scores. Candidates who failed could
retake the exam. Candidates who passed received a certificate that
warranted entry into any Belgian medical university; there was no
further selection on the part of the universities.

**Development and Examination of Alternate SJT Forms**

**Development of alternate SJT forms.** Prior to outlining the
different approaches, it should be noted that pretesting SJT items
on a pilot sample was not possible because all Belgian high school
seniors were potential candidates and pretesting the items on
existing medical students was considered too great a threat to test
security (the central admission exam determines entry to medical
studies in all Belgian universities). It was also legally forbidden to
include nonscored items in the actual tests.

Each year, two alternate SJT forms were developed that were
intended to be comparable. In the early years, the random assign-
ment model was followed. In this strategy, only the domain tar-
geted was considered to be a radical and was therefore held
constant across alternate forms. Thus, both SJT forms measured
interpersonal/communication skills (i.e., skills other than cognitive
ability) related to the interaction between a physician and a patient,
consisting of short descriptions of key interpersonal situations that
physicians are likely to encounter with patients. All other SJT
features were deemed incidentals. Accordingly, the critical inci-
dents, their context, and the exact wording were allowed to vary.
To develop forms according to this first strategy, research assis-
tants interviewed 10 experienced physicians and professors in
general medicine (8 men, 2 women; average age = 39.5 years;
average years of experience = 12.5) to collect critical incidents
related to the domain of interest (i.e., interpersonal/communication
skills related to the interaction between a physician and a patient).
In total, 174 critical incidents were gathered. Research assistants
familiar with SJT development then used these critical incidents to
construct item stems. Next, another group of subject matter experts
was asked to generate response alternatives. Accordingly, a large
number of SJT items was created. These items were then randomly
distributed across SJTs. There was no deliberate attempt to build
the same critical incidents into each SJT. The only requirement
was that the problem situations be representative of the domain.

At the end of 1999, the random assignment approach was
replaced by the incident isomorphic procedure. This meant that
both the domain and the critical incidents (e.g., handling com-
plaints of a patient, conveying bad news) were considered to be
radicals. Thus, these did not vary across forms. All other SJT
features (e.g., the specific context wherein the critical incident and
responses were embedded, linguistic and grammar changes to item
stem and options) were considered to be incidentals. Accordingly,
they were allowed to vary. To put this perspective into practice,
pairs of items that were designed to reflect the same critical
incident (e.g., a physician dealing with a patient who refuses
medication) were developed. Yet, the context wherein the item
stems and options of a pair were embedded was different. One of
each pair was then assigned to alternate forms. Appendix A pre-
seats an item pair of the incident isomorphic procedure. As shown,
the incident (i.e., refusal to take the prescribed medication) is held
constant, whereas the context of item stems and options varies.

After the examinations in 2003, the incident isomorphic ap-
proach was replaced by the item isomorphic approach. Not only
the domain and the critical incidents included, but also the context
wherein they were embedded and the ways of responding to them,
were considered radicals. Accordingly, these aspects did not vary
across forms. That is, pairs of items were developed that reflected
the same critical incident and the same context (item stem and
options). Only wording and grammar variations were allowed
because these aspects were considered to be incidentals. Appendix
B presents an item pair of the item isomorphic approach.

In the end, each SJT form contained 30 questions of the
multiple-choice type, with four response alternatives. The scoring
key was developed in the same way for each SJT form. Each year,
about 10 experienced physicians (professors in general medicine)
independently completed all items. Agreement among the experts
was generally satisfactory (Cohen’s kappas > .70), and discrep-
ancies were resolved upon discussion, leading to the scoring rule.
In some cases, it was necessary to change or remove the items or
options and insert new ones.

**Check of characteristics of SJTs.** We conducted two studies to
check whether the SJT developmental process reflected the char-
acteristics (Table 1) of the three models of constructing alternate
SJT forms. The first study was conducted with five medical
experts (3 men, 2 women; M age = 44.2 years; M years of
experience with patients = 16). These experts were given 20 pairs
of items. These item pairs were randomly chosen from the total
item pool that had been developed across all of the years that the
SJT had been used. The medical experts were asked to rate the
item pairs on the various degrees of similarity as described in
Table 1 (domain similarity, critical incident similarity, context
similarity, and linguistic similarity). Results are presented in Table
2. Planned comparison tests showed that medical experts who had
familiarity with the area being tested rated the similarity of the
item pairs in accordance with the characteristics built into the
various alternate-form development approaches as described in
Table 1. For example, there was no difference between the ap-
proaches in terms of domain similarity. In addition, the random
assignment approach was the only approach that was rated low on
critical incident similarity, whereas the item isomorphic approach
was the only approach that was rated high on context similarity.

The second study was conducted to check whether experts in
test development would be able to distinguish item pairs that had
been developed according to different alternate-form development
approaches. To this end, we asked 5 psychology doctoral students
(2 men, 3 women; M age = 28 years) who had experience with test
and scale development to sort each of the same 20 item pairs into
one of the three alternate-form development categories. Results
showed that experts generally had no problem sorting the item
pairs in the correct alternate-form development category, as the
average kappa was .80 (range = .70–.92), indicating very good
agreement among the experts. Taken together, these two studies confirmed that the strategies used in our main study correctly operationalized the different perspectives in Table 1.

**Criterion Measure**

Criterion data were gathered from students who had passed the exam and had completed the first year of medical studies in one of the universities in Belgium. In particular, we retrieved archival data on students’ scores in interpersonally oriented courses. In these courses, interpersonal and communication skills are taught. Note that these courses are not void of medical subject matter. However, the medical subject content is secondary to the skills of communicating with patients, asking questions, dealing with their complaints, and so on. We gathered interpersonally oriented criterion data because these are especially useful for validating the SJT used (see Lievens et al., 2005a).

To ensure that the criterion had not changed across years, only scores on interpersonally oriented courses that had exactly the same content description and had been taught by the same professors were included. Inspection of the university curricula showed that the curriculum was drastically changed in 1999, as most universities abandoned a conventional medical curriculum with a heavy emphasis on science courses in the first year. Beginning in 1999, interpersonally oriented courses were taught in the first year. Therefore, criterion data for 1998 could not be used.

Empirical evidence that the criterion had not changed across years was given by the fact that the correlation between the interpersonally oriented course grades and overall grade point average (GPA) was similar across the three time periods: \( r = .41 \) for the first time period (random assignment), \( r = .39 \) for the second time period (incident isomorphic), and \( r = .43 \) for the third time period (item isomorphic).

**Results**

**Preliminary Construct Validity and Reliability Analyses**

As noted, it is important that the different approaches to generating items do not produce a change in the construct being measured at the total score level. Hence, we conducted various analyses to investigate the internal consistency and construct validity of the alternate SJT forms. These analyses were conducted on the basis of the data from the alternate forms of the SJT included in this study. First, we examined the correlations of the SJT forms with the cognitive ability test that was also part of the admission exam. This correlation was consistently low (uncorrected rs varying from .08–.11). Second, factor analyses (principal components with varimax rotation) yielded a consistent picture as more than 10 factors (with eigenvalues greater than 1) were typically extracted, explaining more than 50% of the variance. There was no dominant first factor. Third, the internal consistency coefficients of the SJTs across the three approaches were very similar (as \( = .26–.36 \)). This might be due to the fact that our SJTs consisted of only 30 items with a dichotomous scoring scheme (SJT items were scored with either 0 or 1). SJTs typically have more items and use scoring schemes with a wider range of possible scores (see Motowidlo, Dunnette, & Carter, 1990).

Two important conclusions emerged from these results. First, our construct validity and low internal consistency coefficients confirmed prior research and therefore were not unexpected (Chan & Schmitt, 1997; Clause et al., 1998; Schmitt & Chan, 2006). SJTs are methods that enable the measurement of a variety of constructs within a given domain that are difficult to interpret. In this case, interpersonal (communication) skills are a multidimensional construct that can be broken down into various subdimensions such as listening, being empathetic, using appropriate language, or seeking input from patient. Second, these factor analytic and internal consistency results were relatively similar across the SJT forms developed. This consistency provided some important preliminary information in light of our further analyses.

**Alternate-Form Consistency**

To test the first set of hypotheses, we computed correlation coefficients between the SJT forms. As this was a high-stakes selection context, two comments regarding these correlation coefficients are in order. First, these correlation coefficients should not be regarded as test–retest reliability coefficients, because the key condition underlying test–retest reliability studies is violated in a high-stakes context. In fact, test–retest reliability studies are built on the assumption that the person does not change. In addition, in a test–retest reliability study all individuals who take the test the first time also retake the test. Not all of these are the case in high-stakes selection, where only the people who failed the first time are strongly motivated to perform better on the retest. Hence, in this study, we refer to the correlations between alternate forms as **consistency correlations** to avoid confusion with reliability coefficients. Second, it is clear that generally acceptable test–retest

---

**Table 2**

**Summary of Medical Expert Ratings of Item Pairs on Degrees of Similarity**

<table>
<thead>
<tr>
<th>Type of similarity</th>
<th>Random assignment ((k = 8))</th>
<th>Incident isomorphic ((k = 6))</th>
<th>Item isomorphic ((k = 6))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(M)  (SE)</td>
<td>(M)  (SE)</td>
<td>(M)  (SE)</td>
</tr>
<tr>
<td>Domain similarity</td>
<td>4.77s .16</td>
<td>4.50s .27</td>
<td>4.63s .15</td>
</tr>
<tr>
<td>Critical incident similarity</td>
<td>1.87s .21</td>
<td>4.87s .13</td>
<td>5.00s .00</td>
</tr>
<tr>
<td>Context similarity</td>
<td>1.37s .10</td>
<td>1.20b .13</td>
<td>4.63s .11</td>
</tr>
<tr>
<td>Linguistic similarity</td>
<td>1.00b .00</td>
<td>1.43b .28</td>
<td>4.20b .21</td>
</tr>
</tbody>
</table>

Note. \(N = 5, k = \) number of randomly chosen item pairs. Item pairs were rated on a 5-point scale, with higher values indicating higher similarity. Means with different subscripts differed at \(p < .01\).
reliability values (> .80) are too stringent here, as the above illustrates that the alternate-form correlations obtained in a high-stakes context will be lower than the test–retest correlations obtained in a lab context. To assess whether the consistency values for the SJT were acceptable, we compared them to the consistency values obtained for the cognitive ability test that was also used in the high-stakes context of the admission exam. This 50-item cognitive ability test is a measure of general mental ability (GMA) as it consists of items formulated in either verbal, numeric, or figural terms. The consistency values for the cognitive ability tests provide a good benchmark, because parallel forms of this GMA test were developed on the basis of randomly sampling items from an item bank with known item properties (see Nunnally & Bernstein, 1994).

Table 3 presents the SJT consistency correlations broken down by alternate-form development approach. As only students who failed the first time participated in the retest, Table 3 also shows correlations that were corrected for range restriction. To this end, we used indirect range restriction formulas (Thorndike’s Case 3), because candidates were selected on the basis of a third variable (i.e., selected on the basis of a cutoff score determined on an operational composite that was a weighted sum of each of the admission exam tests used). Apart from range restriction corrected values, Table 3 also presents values that took the number of test items into account (50 items for the cognitive ability test and 30 items for the SJT) on the basis of the Spearman Brown prophecy formula.

There was a significant difference between the three alternate-form development approaches. In support of Hypothesis 1a, the consistency among two alternate SJT forms was significantly (z = 7.01, p < .01) higher for the item isomorphic approach (uncorrected r = .54, corrected r = .57) than for the incident isomorphic approach (uncorrected r = .32, corrected r = .41). In line with Hypothesis 1b, the consistency among two alternate SJT forms was significantly (z = 8.31, p < .01) higher for the incident isomorphic approach (uncorrected r = .32, corrected r = .41) than for the random assignment approach (uncorrected r = .21, corrected r = .22).

As could be expected, the consistency values of the GMA test parallel forms did not differ significantly across time, with values of .53 (corrected r = .62), .55 (corrected r = .70), and .54 (corrected r = .67). This confirmed that the GMA test was a useful benchmark against which to assess the degree of consistency in this high-stakes selection context. Table 3 also shows that only SJT forms developed on the basis of the item isomorphic approach yielded consistency values (.68) that were almost equal to those of the GMA (.67), which seems adequate in light of the high-stakes testing context.

On the basis of the Spearman Brown prophecy formula, we also examined how many SJT items developed on the basis of the random assignment or incident isomorphic approaches were needed to obtain the consistency evidenced by the item isomorphic approach. These analyses revealed that an SJT developed on the basis of the random assignment approach needed 5 times the number of items of an SJT developed according to the item isomorphic model to obtain the same alternate-form consistency. For SJTs constructed according to the incident isomorphic approach, 2.8 times the number of items was needed.

Finally, in ancillary analyses we checked whether the change in presentation format (video vs. written) influenced the alternate-form consistency results. This was done by comparing data of the last 2 years of the incident isomorphic approach. When original and alternate forms of a video-based SJT were used, the uncorrected correlation was .58, whereas this correlation was .57 for original and alternate versions of a written SJT. Thus, the change in presentation format did not affect the alternate-form consistency results.

Retest Effects

To test the set of hypotheses about retest effects (mean score changes upon retesting), we computed d values. These d values were obtained by subtracting the score on the first examination from the score on the second examination, and dividing by the pooled standard deviation (Hunter & Schmidt, 2004, p. 277, Equa-

Table 3
Consistency Correlations of SJT Forms Broken Down by Alternate-Form Development Approach

<table>
<thead>
<tr>
<th>Development approach</th>
<th>Consistency index</th>
<th>Corrected for indirect range restriction</th>
<th>Corrected for number of GMA items (50)</th>
<th>Corrected for number of SJT items (30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SJT</td>
<td>Observed</td>
<td>.21,</td>
<td>.22</td>
<td>.34</td>
</tr>
<tr>
<td>GMA</td>
<td></td>
<td>.53,</td>
<td>.62</td>
<td>.62</td>
</tr>
<tr>
<td>Incident isomorphic (n = 1,385)</td>
<td></td>
<td>.32,</td>
<td>.41</td>
<td>.56</td>
</tr>
<tr>
<td>SJT</td>
<td></td>
<td>.55,</td>
<td>.70</td>
<td>.70</td>
</tr>
<tr>
<td>GMA</td>
<td></td>
<td>.54,</td>
<td>.57</td>
<td>.68</td>
</tr>
</tbody>
</table>

Note. Correlations with different superscripts differed at p < .01. SJT = situational judgment test; GMA = general mental ability.
tion 7.4). Positive effect sizes meant that the second examination score was higher than the first one.

Table 4 presents the SJT retest effects broken down by alternate-form development approach. There was mixed support for our hypotheses. Hypothesis 2b was confirmed, as retest effects were larger for the item isomorphic approach \((d = .67)\) than for the random assignment approach \((d = -.14)\). The negative \(d\) for the random assignment approach can be understood from the notion that the alternate SJT is essentially a different SJT. However, there was no support for Hypothesis 2a. Retest effects were not larger for the item isomorphic approach \((d = .27)\) than for the incident isomorphic approach \((d = .67)\). Note that the reliability (internal consistency) of the SJTs developed according to either one of the three approaches did not differ from one another (see above). Hence, differences in reliability are not an explanation for these results.

Criterion-Related Validity

Finally, we examined the criterion-related validity of the different SJT forms. To this end, we compared the correlation between individuals’ original test score and the criterion with the correlation between individuals’ repeat test score and the criterion. These analyses were based on a relatively small sample size. Hence, the power to find statistically significant differences between the approaches was low. The small sample size was due to the fact that validity coefficients could be computed only on the basis of people who had participated in two administrations, passed on their second attempt, started medical education, and successfully completed the first year of schooling. In addition, interpersonal GPA (see criterion description) was the most conceptually relevant criterion for the SJT. However, interpersonal courses were not taught in all medical schools in the first year, further limiting the sample size in these analyses.

The uncorrected and corrected (for indirect range restriction) validity coefficients are presented in Table 5. The validity coefficients obtained varied from .11 to .18. Although these validity coefficients might be regarded as relatively low, it should be noted that they were based on first-year grades. Along these lines, Lievens et al. (2005a) found similar validity coefficients for the first year, with the validity of the SJT gradually increasing in later years as interpersonal courses gained in importance in determining GPA.

Table 4
Re-test Effects Associated With SJT Broken down by Alternate-Form Development Approach

<table>
<thead>
<tr>
<th>Development approach</th>
<th>Initial test</th>
<th>Retest</th>
</tr>
</thead>
<tbody>
<tr>
<td>SJT ((n = 703))</td>
<td>19.55 3.14</td>
<td>19.10 3.24</td>
</tr>
<tr>
<td>SJT ((n = 1,385))</td>
<td>17.93 2.97</td>
<td>19.90 2.97</td>
</tr>
<tr>
<td>SJT ((n = 1,273))</td>
<td>17.68 3.01</td>
<td>18.51 3.03</td>
</tr>
</tbody>
</table>

Note. The \(d\) values are effect sizes computed by \((M_{\text{second}} - M_{\text{first}})/\text{pooled } SD\). Positive effect sizes mean that the second examination score was higher than the first one. SJT = situational judgment test.

We used Meng, Rosenthal, and Rubin’s (1992) \(Z\) test of the difference between dependent correlations to examine whether the validity for the initial test differed significantly from the validity for the retest. For none of the approaches was the validity for the initial test significantly different from the validity for the retest.

Discussion

Theoretical Contributions

At a theoretical level, this study provides a first important step in discovering how similar an SJT needs to be to provide consistent measurement across alternate forms. Specifically, we used the radicals—incidents approach from item generation theory to identify three different perspectives about what might determine structural features in SJT items. One perspective posited that the major determining feature of SJTs is the domain to be targeted. Empirical results did not provide support for this perspective. The poor alternate-form consistency results of the random assignment approach revealed that the domain sampled by an SJT should not be considered to be the only radical (i.e., a major determining characteristic) of SJTs. An interpersonal SJT correlated only in the .20s with an alternate interpersonal SJT (with items being randomly assigned to either one of these SJTs). It should be noted, though, that this correlation was obtained for two SJT forms, each with 30 items. Better results for the random assignment approach are to be expected when SJTs consist of more items. In any case, the poor performance of the random assignment approach is not a trivial result. For example, it would be surprising to find that an original form of, for instance, a personality scale correlates in the .20s with an alternate form that targets the same domain (with items being randomly assigned to either one of these personality scale forms; Barrick & Mount, 1993).

The second perspective posited that the domain to be targeted and the specific critical incidents gathered to develop SJT items are radicals. According to this perspective, all other SJT features (e.g., the specific context wherein the critical incident and options are embedded, linguistic and grammar changes to item stem and options) are incidentals. There was also no support for this perspective. Although the incident isomorphic strategy obtained significantly higher consistency than the random assignment strategy, it was still below acceptable standards. Hence, a key finding is that conceptualizing the domain and critical incident gathered as radicals is a necessary albeit insufficient condition to obtaining ac-
ceptable alternate-form consistency. Apparently, there are other features (radicals) that might make SJTs substantially different.

Along these lines, the third perspective posited that the context wherein the critical incident and responses are embedded is also a radical. According to this perspective, incidentals in SJT items might refer only to the linguistic and grammar changes to item stems and options. We found empirical support for this perspective of conceptualizing radicals and incidentals in SJTs. The consistency obtained for the item isomorphic approach was equal to the one obtained for a cognitive ability test developed according to well-established domain-sampling procedures. This confirms that the context wherein the critical incident and the item responses are embedded should also be considered a radical in SJTs. From a conceptual point of view, an important point in these alternate-form consistency results is that apparently even slight variations in the problem situations included across SJTs might result in lack of consistency. Therefore, acceptable consistency across alternate forms of such predictors seems to be expected only when contextual factors are held constant (are considered radicals) across forms. This specificity issue is not unique to SJTs. For example, exercise specificity is a well-known phenomenon in assessment centers (Bramnick, Michaels, & Baker, 1989; Lance, Lambert, Gewin, Lievens, & Conway, 2004). In addition, context and task specificity have been shown to determine performance-based tests in a variety of other areas, including hands-on science tasks (Shavelson et al., 1991), bar examinations (Sackett, 1998), military examinations (Shavelson, Mayberry, Li, & Webb, 1990), and the assessment of clinical skills using standardized patients (Swanson, Norman, & Linn, 1995; Vu & Barrows, 1994).

Consistent with the assumptions underlying item generation theory, none of the alternate-form development approaches provided a shift in the constructs being measured at the total score level. As shown by the correlations between the SJTs and cognitive ability, the SJT (regardless of the alternate-form development approach) always correlated poorly with cognitive ability. In addition, we found no effects of the alternate-form development approaches on criterion-related validity. This might be explained by the fact that all SJTs were designed to capture the same domain and included critical incidents relevant to this domain. Accordingly, the conceptual overlap between predictor and criterion was ensured in each of the approaches (Ryan & Greguras, 1998).

In short, the diverging alternate consistency results underscore the importance of carefully identifying the radicals and incidentals in SJTs. If an incidental turns out to be an important determining feature (i.e., radical), consistency is bound to suffer, as was demonstrated by the random assignment and incident isomorphic approaches. Such a result might undermine the psychometric defensibility of the SJT. More broadly, this also calls for a theory about how people answer SJTs (Motowidlo, Hooper, & Jackson, 2006) or a taxonomy about what makes some SJT items more or less difficult than others.

**Practical Contributions**

Public or private organizations with testing programs often need to develop alternate forms. Otherwise, the testing program runs the risk of being compromised. Figure 1 brings together the different alternate-form development strategies that might be followed. In addition, key issues that differentiate the various approaches are listed along with contextual factors. The first question is whether item pretesting is possible. If pretesting items is possible (e.g., by administering items on a pilot group or by embedding them in existing tests), test developers might use relatively established procedures. In case of tests with multidimensional poorly understood constructs (e.g., SJTs), one might use the procedure that Oswald, Friede, Schmitt, Kim, and Ramsay (2005) recently developed (Situation 1). In the case of tests with relatively unidimensional well-understood constructs (e.g., cognitive ability tests, knowledge tests), constructing parallel tests is straightforward, as there exists a vast literature on the use of domain-sampling (Nunnally & Bernstein, 1994; Situation 2) or item generation approaches (in the case that an item generation scheme exists; Situation 3).

This study has especially practical implications for situations wherein pretesting is considered too great a threat to test security. When organizations face a situation wherein item trialing is not possible, the issue they need to attend to is whether the test measures unidimensional well-understood constructs or multidimensional poorly understood constructs. Although item generation theory has been frequently used for tests measuring unidimensional well-understood constructs (Irvine & Kylomen, 2002; Situation 3), this study provides some research-based guidelines for generating alternate forms of a multidimensional test with poorly understood constructs such as an SJT (Situations 4 or 5). So far, test developers have typically been left in the dark with regard to alternate-form construction approaches for these situations. The practical message of this article is not simple; it depends on what test developers value. If one is concerned about possible retest effects (Situation 4), we believe the random assignment strategy might be the best option provided that one is able to use SJTs with a large number of items. If retest effects are not considered to be problematic (Situation 5), we recommend the item isomorphic approach.

Figure 1 also points to temporal and situational factors that might influence the alternate-form development strategy adopted. Temporal factors include the length of the time interval between test administrations and the period of time that the test program has existed. In addition, a situational factor such as the presence of a coaching industry might impact on the feasibility of some approaches. For example, once firms in the test preparation business know the item generation scheme being applied, it becomes easy to prepare applicants for the exam. This is especially the case if the admissions exam consists of a small number of items. The high-stakes setting of this study precluded us from disentangling practice and coaching effects. Therefore, we concentrated on the broader phenomenon of retest effects. Future lab studies should examine whether the different alternate-form development approaches are robust to deliberate coaching (Cullen, Sackett, & Lievens, 2006).

Apart from these general strategies and decisions, this study has some specific guidelines for creating alternate forms of SJTs. First, it is crucial to think in advance carefully about what are radicals and incidentals in SJTs and make informed decisions about this issue. To derive such an item generation scheme, input from subject matter experts seems invaluable. Second, one should keep in mind that small changes in the situations presented in SJTs might affect the alternate-form consistency. Third, our results suggest that it is beneficial to place more structure and constraints...
on the alternate-form development process (i.e., keeping more features constant). This might mean that the context of the item stems and options is held constant. As a final backdrop option for increasing alternate-form consistency, it seems advisable to ensure that the SJT has a large number of items.

Limitations

This study is not without limitations. First, the different types of alternate forms were administered in different years. Ideally, in an experimental research design we would have randomly divided an initial pool of items into thirds, developed items of each part according to one of the three approaches, and administered them to the same sample. This study was conducted in the field in a high-stakes setting, precluding such experimental control. However, it is important to note that there was neither substantial change in the admission exam procedure nor in the make-up of the student samples (the latter were essentially populations). This makes it highly unlikely that our results were due to procedural or sample changes that might have occurred through the years.

A second limitation is related to the generalizability of our results. Our study dealt with an SJT that measured interpersonal skills and that had a dichotomous scoring scheme. Future research should examine whether our results generalize to other SJTs. The study was also conducted in an educational high-stakes context in Belgium. There are some differences between admission practices in Belgium and the United States. For example, in Belgium the admission process is centralized, and the level of selectivity (30% passing rate) is less stringent than in the United States. Yet there are also many similarities in admission practices between Belgium and the United States. For example, in both countries it has been suggested that cognitive predictors be supplemented with SJTs (Lievens et al., 2005a; Oswald et al., 2004). Hence, we believe that our results might generalize to high-stakes testing in other countries such as the United States. In addition, our results might be relevant for high-stakes employment testing in the public sector (e.g., selection of firefighters, law enforcement personnel), as the retesting policies for those jobs share many parallels with retesting policies in high-stakes educational testing. In any case, future studies are needed to examine the generalizability of our results to other settings and countries.

Third, it should be acknowledged that GPA served as the criterion. Therefore, future research should examine whether our re-

Figure 1. Decision tree with recommendations for constructing alternate forms of tests in various retest scenarios.
sults generalize to employment settings with job performance as the criterion. It is worth noting, however, that much validation work in employment settings is against training criteria. Also, while there are certainly differences between work and education settings, note that we find the SJT useful in less academic settings, namely interpersonal courses. It can be argued that activities in such courses are more similar to those at work than are courses in the sciences. In addition, grades on interpersonal courses are often not based on regular exams but on practical exercises.

Implications for Future Research

One of the lessons learned from this study is that it is of key importance to explicitly consider specific structural SJT features (e.g., the critical incidents gathered, the context wherein these incidents are embedded) if one wants to successfully create alternate SJT forms. This puts the attention on item specifications underlying SJTs rather than on the SJT test as a whole. This line of thought might not only be applied to creating alternate SJT forms. It might also be fruitfully used to better understand factors that affect the validity and adverse impact of SJTs. So far, SJT development has been mostly a bottom-up process. That is, subject matter experts provide the critical incidents and potential responses that are then edited and converted by SJT developers into item stems and response alternatives, respectively. However, the universe of SJT items is defined by various factors that researchers are only beginning to understand. Examples are the complexity of the incidents (the number of problems in one SJT item), the presentation format, and response instructions (knowledge-based vs. behavior-oriented). The field needs more studies that vary specific SJT features to understand their effects on relevant variables such as validity and adverse impact. At a practical level, such studies might provide item design rules that inform SJT development practice.

As a second avenue for future research, the applicability of the proposed alternate-form development approaches should be studied for other multidimensional tests such as assessment center exercises. For instance, practitioners might try out an isomorphic approach for constructing alternate forms of the problem descriptions used in case analyses, in-baskets, presentations, role plays, and group discussions. The application of isomorphic approaches to situational interviews is another possibility.

Third, future studies should examine applicants’ reactions to item isomorphic approaches. We expect that applicants might react somewhat negatively to item isomorphic alternate-form development approaches. Applicants might perceive them to be low on a procedural justice dimension, such as “opportunity to perform,” as they are given the same test apart from some linguistic changes. They might not feel that they are really given another chance to retake the test and show their capabilities.

References


review of practice and constructs assessed. International Journal of Selection and Assessment, 9, 103–113.

Appendix A

Example of Incident Isomorphic Approach

Original item:

Patient: So, this physiotherapy is really going to help me?
Physician: Absolutely, even though the first days it might still be painful.
Patient: Yes, I suppose it will take a while before it starts working. Physician: That is why I am going to prescribe a painkiller. You should take 3 painkillers per day.
Patient: Do I really have to take them? I have already tried a few things. First, they didn’t help me. And second, I’m actually opposed to taking any medication. I’d rather not take them. They are not good for my health.
What is the best way for you (as a physician) to react to this patient’s refusal to take the prescribed medication?

a. Ask her if she knows something else to relieve the pain.
b. Give her the scientific evidence as to why painkillers will help.
c. Agree not to take them now, but also stress the importance of the physiotherapy.

d. Tell her that, in her own interest, she will have to start changing her attitude.

Alternate item:

Physician: I am going to prescribe some medication that should substantially improve the symptoms in the next days or so.
Patient: Hopefully. Yes.
Physician: I will prescribe antibiotics.
Patient: Antibiotics? I’m actually opposed to taking antibiotics. People say they are no good. I’d rather not take them.
What is the best way for you (as a physician) to react to this patient’s refusal to take the prescribed medication?

a. Tell him that, in his own interest, it is important that he take the antibiotics.
b. Clarify in a friendly way that such an attitude will not solve his problems.
c. Explain that all scientific experts agree that antibiotics are needed here.
d. Emphasize that his problems will not go away without antibiotics.
Appendix B

Example of Item Isomorphic Approach

Original item:

*Patient:* It is really aching.

*Physician:* Do you have an idea yourself about what could be the cause?

*Patient:* I don’t know. Perhaps it is family-related, because I know a lot of family members have a back problem. Does this mean it is genetic?

*Physician:* Perhaps.

*Patient:* In fact, my nephew. He suffers very often from back pain. Sometimes I hear he is okay, but he also has to lie in bed. What is the best way for you (as a physician) to continue the conversation?

a. Clarify that the information regarding her family member is not relevant here.

b. Inquire whether back pain is prevalent among other family members.

c. Ask whether her nephew is Frank Rice, who is also one of your patients.

d. Ask more direct questions about the back pain of the patient.

Alternate item:

*Patient:* It is actually painful.

*Physician:* Do you have a clue yourself what caused your back pain?

*Patient:* I don’t have a clue. Perhaps it runs in the family, because a lot of people in my family suffer from back pain. However, does this mean that it is heritable or something like that?

*Physician:* Could be.

*Patient:* My nephew, for instance, very often has back problems. It comes and goes. Sometimes it is okay, but on other occasions he has to stay home. What is the best way for you (as a physician) to go on with the conversation?

a. Make clear that the details about his nephew are not important.

b. Find out whether back problems are also prevalent among another member of his family.

b. Tell him that his nephew, Joseph Dune, is also one of your patients.

d. Ask more straightforward questions about the patient’s back problems.