UNIVERSITÄT OSNABRÜCK

Situational Judgment Tests

Untersuchungen zu low-fidelity Simulationen unter besonderer Berücksichtigung grundlegender psychometrischer Eigenschaften

Dissertation zur Erlangung des Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
eingereicht am Fachbereich Humanwissenschaften
der Universität Osnabrück

vorgelegt von

Dipl.-Psych. Nadine Kasten

aus Minden/Westf.

Osnabrück, 2017

Erstprüfer: Prof. Dr. Thomas Staufenbiel

Zweitprüfer: Prof. Dr. Karsten Müller

Danksagung

Im Rückblick auf die vergangenen Jahre möchte ich mich bei vielen Menschen bedanken, die zum Abschluss meiner Dissertation beigetragen haben. Ein ganz besonderer Dank gilt meinem Doktorvater Thomas Staufenbiel, ohne dessen Anregung ich nie den Weg in die Wissenschaft gefunden hätte. Auch deine stets objektive und differenzierte Meinung zum BVB waren für mich während meiner Zeit in Osnabrück stets ein Quell der Freude.

Auch bei Karsten Müller möchte ich mich für die Bereitschaft bedanken, die Promotion zu betreuen.

Alexander Freund gilt großer Dank. Ohne deine hilfreichen Impulse, deinen Zuspruch und Unterstützung, wäre dieses Projekt nicht zum Abschluss gekommen.

Ich danke meinen Kollegen und Freunden Julia Hülsmann, Lisa und Dirk Zimmermann, Kathrin Wunsch, Ramona Wurst, Timo Gnambs, Christopher Klanke, den Bewohnern der Roopkommune, Jennifer und Benjamin Molitor, Łukasz Stasielowicz, Judith Rickers, Swarley Spieß, Thomas Seppelfricke und Rohangis Mohseni für viele Stunden der Ablenkung von der Arbeit. Gleiches gilt für Sarah Konerding, die es zudem vermochte, auch bei der fünften Erklärung zum adäquaten Umgang mit Reisekostenerstattungsformularen eine stoische Geduld an den Tag zu legen. Darüber hinaus danke ich meiner Familie, die mich immer unterstützt hat.

Der größte Dank gilt Oli. Für alles.

Hinweise zur Dissertation

Bei der vorliegenden Arbeit handelt es sich um eine kumulative Dissertation gemäß §10 Absatz (3) der aktuellen Promotionsordnung des Faches Psychologie. Die drei Artikel wurden bei Zeitschriften mit peer-review Verfahren eingereicht bzw. sind bereits veröffentlicht. Teilergebnisse der Arbeit wurden im Rahmen folgender Kongressvorträge präsentiert:

Kasten, N. & Freund, P. A. (2013, September). *Eine metaanalytische Reliabilitätsgeneralisierung von Situational Judgment Tests (SJTs).* Vortrag auf der 12. Arbeitstagung der Fachgruppe Differentielle Psychologie, Persönlichkeitspsychologie, und psychologische Diagnostik in Greifswald.

Kasten, N. & Staufenbiel, T. (2015, Mai). *A construct-oriented development approach of situational judgment tests (SJTs).* Vortrag auf der 17. Tagung der European Association of Work and Organizational Psychology (EAWOP) in Oslo, Norwegen.

Kasten, N. & Staufenbiel, T. (2015, September). *Ein konstruktorientierter Ansatz zur Entwicklung von Situational Judgment Tests (SJTs).* Vortrag auf der 9. Fachgruppentagung Arbeits-, Organisations- und Wirtschaftspsychologie (AOW) in Mainz.

Kasten, N., Freund, P. A. & Staufenbiel, T. (2016, September). *„Sweet little lies“: Fakinganfälligkeit konstruktorientierter Situational Judgment Tests im Vergleich zu Persönlichkeitsfragebogen.* Vortrag auf dem 50. Kongress der deutschen Gesellschaft für Psychologie (DGPs) in Leipzig.

Zusammenfassung

Situational Judgment Tests (SJTs) werden den simulationsorientierten Verfahren zugerechnet und werden vorrangig im Rahmen der Personalauswahl eingesetzt. Im Gegensatz zu anderen Simulationen (wie z.B. Arbeitsproben oder Assessment Center) wählen sie einen *low-fidelity* Ansatz, da hier die Reaktionen der Bewerber nicht über tatsächliches Verhalten abgebildet werden. Vielmehr werden Bewerber mit hypothetischen Situationen konfrontiert und aufgefordert aus vorgegebenen Handlungsalternativen diejenige auszuwählen, die in der gegebenen Situation am angemessensten erscheint. Innerhalb der letzten 20 Jahre haben sich SJTs zu populären Verfahren innerhalb Personalauswahl entwickelt. Trotz der hohen Forschungs- und Publikationsdichte bestehen dennoch erhebliche Probleme, wenn es um den Nachweis grundlegender psychometrischer Eigenschaften von SJTs geht. Dementsprechend sollen die vorliegenden Studien vorrangig dazu dienen, unser Verständnis von Reliabilität, Konstruktvalidität und Verfälschbarkeit von SJTs zu erweitern. Im Rahmen von Studie 1 konnte über die metaanalytische Technik der Reliabilitätsgeneralisierung ein mittlerer Wert für das Ausmaß des Messfehlers bei der Verwendung von SJTs bestimmt werden. Entgegen der bisherigen Vermutung, wiesen viele SJTs in den relevanten Primärstudien ein ausreichendes Ausmaß an Messgenauigkeit auf. Zudem konnte der teilweise bedeutsame Einfluss vieler Test- und Studiencharakteristika dargestellt werden. In Studie 2 wurde der Einfluss des Entwicklungsparadigmas auf die Konstruktvalidität von SJTs dargestellt. Über die Anpassung des klassischen Vorgehens zu einem konstrukt-orientierten Ansatz konnte ein konstruktvalides SJT zur Messung von Extraversion, Verträglichkeit und Gewissenhaftigkeit entwickelt werden. Im Rahmen von Artikel 3 werden die Ergebnisse von zwei Studien zusammengefasst,

die sich mit Unterschieden der Fälschungsanfälligkeit dieses SJTs im Vergleich zu einem klassischen Selbstbeurteilungsfragebogen (NEO-FFI) auseinandersetzen. In beiden Studien zeigen sich deutliche Unterschiede, in dem Sinne, dass das SJT weniger Faking aufweist als der NEO-FFI.

Die Ergebnisse der einzelnen Studien, die im Rahmen der Dissertation durchgeführt wurden, adressieren wichtige Fragestellungen, die in der bisherigen Literatur weitestgehend vernachlässigt wurden. Die praktischen und theoretischen Implikation sowie die Limitationen der einzelnen Studien werden diskutiert. Die Synopsis schließt mit einem Ausblick auf Forschungsfragen, die offen geblieben sind und dementsprechend den Ausgangspunkt für weitere Studien darstellen.

Summary

Situational Judgment Tests (SJTs) are simulation-based measurement methods originally designed for the application in personnel selection settings. Unlike other simulation-based methods (e.g. work samples or roleplays), SJTs apply a low-fidelity approach, as applicants are not expected to perform actual behavior. Instead they are presented with a hypothetical work-situation and have to choose an appropiate reaction from a set of different alternatives. Although SJTs have become increasingly popular over the last two decades in both research and practice, they are often only poorly understood on the level of essential psychometric properties. Therefore, the present studies primary aim to foster our understanding of SJT reliability, construct validity, and fakability. In study 1, we applied reliability generalization technique to examine average α coefficients associated with the use of SJTs. Contrary to the predominant negative view in SJT literature, many studies indicated an appropriate level of measurement accurarcy for the applied SJTs. Additionally, moderator analysis substantiated great influence of many test and study characteristics on measurement error. In study 2, we examined the possibillities to target homogeneous constructs using the SJT paradigm. Through the implementation of a construct-oriented way to develop and score SJT items we were able to develop a valid measure of extraversion, agreeableness, and conscientiousness. Article 3 examines the utility of the newly constructed SJT as an alternative measure of personality. Therefore, two studies examine the extent to which it is susceptible towards faking compared to a traditional self-report measure of personality (i.e. NEO-FFI). In both studies standardized mean differences between honest and faked conscientiousness scores were less pronounced for the SJT than for the NEO-FFI. Referring to antecedents of faking

behavior and differences in faking styles, the present studies attempt to further analyze these differences.

The studies included in the present dissertation address important points and thus make a timely contribution to the growing field of SJT research. Implications for practitioners and researchers are discussed as well as limitation of the present studies. The synopsis concludes with a brief discussion of open questions.

Inhaltsverzeichnis

1 Einleitung

Die Auswahl von Mitarbeitern[1] gehört zu den wichtigsten Entscheidungen eines Unternehmens (Görlich & Schuler, 2014; Schuler, 2009). So hängt nicht nur die Produktivität und Effizienz des Unternehmens in entscheidender Weise von der Passung zwischen Mitarbeiter und Stelle ab, auch können sich Fehlentscheidungen sowohl für das Unternehmen als auch für die Mitarbeiter in unterschiedlichen und weitreichenden negativen Konsequenzen manifestieren. Auf Unternehmensseite vor allem der ökonomische Schaden zu bedenken, der mit der Auswahl ungeeigneter Mitarbeiter verbunden ist. Im Rahmen von Auswahlprozessen besteht daher das Ziel psychologischer Diagnostik primär darin, aus einem Pool von Bewerbern diejenigen auszuwählen, die am besten für die ausgeschriebenen Stellen geeignet sind. Obwohl Eignung ein vielschichtiges Konzept darstellt, definiert es sich aus Organisationssicht vorrangig über den Zusammenhang der diagnostischen Aussagen mit der gegenwärtigen und zukünftigen beruflichen Leistung (Marcus, 2011). Die Suche nach validen Prädiktoren von Leistungskriterien ist folglich seit jeher fester Bestandteil der Eignungsdiagnostik (Schuler, 2014; Weuster, 2008). Die Beurteilung der Brauchbarkeit eines Auswahlinstruments bemisst sich allerdings nicht ausschließlich an dessen Vorhersageleistung. In den letzten beiden Jahrzehnten haben einschneidende gesellschaftliche und politische Veränderungen zu einer Erweiterung des Beurteilungsspektrums um weitere Qualitätsaspekte geführt. So sind beispielsweise Aspekte der Testfairness zunehmend rechtlich fixiert, z.B. in Form von

[1] Zugunsten der besseren Lesbarkeit wird durchgängig die männliche Form verwendet. Selbstverständlich sind jedoch stets beide Geschlechter gemeint.

Antidiskriminierung- und Gleichstellungsgesetzen (Marcus, 2011). Darüber hinaus haben der demographische Wandel und der zunehmende Fachkräftemangel zu einem intensiveren Wettbewerb um hochqualifizierte Mitarbeiter geführt. Dieser sogenannte „*War for Talents*“ (Kirchgeorg & Müller, 2013, S. 73) hat in großem Maße dazu beigetragen, dass neben der organisatorischen Position auch zunehmend die individuelle Bewerberperspektive im Auswahlverfahren betont wird (Achouri, 2010). Als besonders wichtig hat sich in diesem Zusammenhang die Wahrnehmung und die Akzeptanz der eignungsdiagnostischen Situation seitens der Bewerber erwiesen (Gilliland, 1993; 1994; Schmitt & Gilliland, 1992; Schuler, 2009). Diese subjektiven Bewerberreaktionen hängen konzeptuell von einer Reihe von Einflussgrößen ab, darunter auch die Art der eingesetzten Auswahlverfahren. Insbesondere für deren Tätigkeitsbezug konnte in vielen Studien ein positiver Einfluss auf die wahrgenommene Akzeptanz der Auswahlsituation seitens der Bewerber nachgewiesen werden (Bauer, Truxillo, Sanchez, Craig, Ferrara & Campion, 2001; Chapman, Uggerslev, Carroll, Piasentin & Jones, 2005). Die Berücksichtigung der Frage, wie der Bewerber auf die Auswahlsituation reagiert und durch welche Faktoren diese Wahrnehmung beeinflusst wird, ist nicht ausschließlich durch ein grundlegendes Forschungsinteresse gespeist, sondern weist durchaus praktischen Nutzen für die Organisationen auf (Kanning, 2015). So deuten empirische Untersuchungen darauf hin, dass diese bedeutsame Auswirkungen sowohl auf die Motivation und die Leistung der Bewerber (McCarthy, van Iddekinge, Lievens, Kung, Sinar & Campion, 2013) als auch auf die Reaktion hinsichtlich einer Stellenofferte (Carless, 2003) und die anschließende Bewertung und Einstellung gegenüber dem Unternehmen (Hausknecht, Day & Thomas, 2004) hat.

Zusammenfassend lässt sich also festhalten, dass von einem eignungsdiagnostischen Verfahren demnach nicht nur eine angemessene Prognoseleistung erwartet wird, es soll sich ebenfalls durch Akzeptanz seitens der Bewerber, wahrgenommene und tatsächliche Fairness auszeichnen. In diesem Zusammenhang werden zunehmend *Situational Judgment Tests* (SJTs) als eignungsdiagnostische Repräsentanten dieser positiven Merkmalskombination genannt (Landy, 2007). SJTs sind ‚neuartige‘ Prädiktoren der Personalauswahl, die vor allem

in den letzten 25 Jahren stark an Verbreitung gewonnen haben, deren historischer Hintergrund aber bis in die 1920er zurückreicht (Behrmann, 2007). Insbesondere der Artikel von Motowidlo, Dunnette und Carter (1990) bereitete durch die Definition von allgemeinen Konstruktions- und Anwendungsprinzipien den Weg für den Vormarsch von SJTs. Gemessen an der Publikationsdichte und Anwendungshäufigkeit, haben SJTs einen festen Platz sowohl in der Forschung als auch in der Praxis eingenommen (Campion, Ployhart & MacKenzie, 2014). Um zu verdeutlichen, wie SJTs typischerweise aufgebaut sind, wird in Abbildung 1 ein Beispielitem präsentiert. Wie abgebildet besteht der Itemstamm von SJT Items aus der Darstellung berufsrelevanter, erfolgskritischer und häufig problematischer Situationen. Dieser Situationsbeschreibung folgt ein geschlossenes Antwortformat. Das bedeutet, dass die Handlungsreaktion auf die dargestellte Situation nicht frei vom Bewerber selbst generiert wird, sondern vielmehr verschiedene Handlungsoptionen präsentiert werden. Diese sollen dann gemäß der spezifischen Instruktion vom Bewerber evaluiert werden, hier z.B. anhand der wahrgenommenen Effektivität.

Vor allem im Rahmen von Personalauswahlverfahren ist der Einsatz von SJTs zunehmend gebräuchlich (Chan & Schmitt, 2005), aber auch darüber hinaus finden sie verstärkt Anwendung in verschiedensten Kontexten zur Vorhersage oder Messung unterschiedlicher Konstrukte, z.B. bei der Erfassung von Teamleistungen (Stevens & Campion, 1999; Wang, MacCann, Zhuang, Liu & Roberts, 2009), Studierfähigkeit (Oswald, Schmitt, Kim, Ramsay & Gillespie, 2004), Verhandlungskompetenz (Behrmann, 2011), bei der Erfassung interpersoneller Fähigkeiten (Lievens, 2013), oder der Evaluation von Trainingsprogrammen (Hauenstein, Findlay & McDonald, 2010). Allerdings hat diese verstärkte Anwendung von SJTs nicht zu einem verbesserten Verständnis auf Ebene der theoretischen und konzeptuellen Vorstellungen geführt. Ganz im Gegenteil erscheint es eher so, dass sich mit steigender Publikationsdichte und Anwendungshäufigkeit auch zunehmend unterschiedliche Vorstellungen darüber manifestieren, was SJTs eigentlich genau definiert bzw. welche Charakteristika für SJTs essentiell sind. Zudem weisen SJTs in der Regel noch erhebliche Problematiken bezogen auf den Nachweis grundlegender psychometrischer Eigenschaften auf. Die vorliegenden

You assigned a very high profile project to one of your project managers. During each of the project update meetings, your project manager indicates that everything is going as scheduled. Now, one week before the project is due, your project manager informs you that the project is less than 50 % complete.

Given the present situation, rate the effectiveness of the following reactions on a scale ranging from 1 (= extremely bad) to 7 (= extremely good).

Personally take over the project and meet with the costumer to determine critical requirements.	①	②	③	④	⑤	⑥	⑦
Meet with the costumer to extend the deadline. Talk to the project manager about how his behavior has jeopardized the relationship with the costumer.	①	②	③	④	⑤	⑥	⑦
Fire the project manager and take over the project yourself.	①	②	③	④	⑤	⑥	⑦
Coach the project manager on how to handle the project more efficiently.	①	②	③	④	⑤	⑥	⑦
Do not assign any high profile jobs to this project manager in the future.	①	②	③	④	⑤	⑥	⑦

Abbildung 1. Beispielitem nach McDaniel & Whetzel (2007)

Artikel nehmen diese Unzulänglichkeiten als Ausgangspunkt. Dementsprechend liegen die Schwerpunkte der empirischen Arbeiten auf der Messgenauigkeit von SJTs (Artikel 1), dem Nachweis von Konstruktvalidität (Artikel 2) und der Quantifizierung der Verfälschbarkeit von SJTs (Artikel 3).

Die Synopsis gliedert sich in drei Teile: Der erste Teil umfasst den theoretischen und empirischen Hintergrund zu SJTs. Dazu wird zunächst ein differenzierter Überblick über definitorische Aspekte von SJTs und den charakteristischen Aufbau von SJT Items gegeben (Kapitel 1.1). Anschließend soll die bestehende Heterogenität zum Ausdruck kommen, die sich in der Literatur findet, wenn es um die Darstellung bedeutsamer Testcharakteristika von SJTs geht (Kapitel 1.2). Schließlich werden die psychometrischen Eigenschaften von SJTs umrissen, wobei ein Schwerpunkt auf die Aspekte gelegt wird, die Ausgangspunkte der vorliegenden empirischen Arbeiten darstellen (Kapitel 1.3).

Der zweite Teil der Synopsis (Kapitel 2) beinhaltet die Darstellung der empirischen Arbeiten, die im Rahmen der Dissertation durchgeführt wurden, und die bei internationalen Journals eingereicht bzw. bereits veröffentlicht wurden. Der erste Artikel (Kapitel 2.1) beschäftigt sich mit der Reliabilität von SJTs. Über die metaanalytische Methodik der Reliabilitätsgeneralisierung bestehen die primären Ziele dieses Artikels darin, eine Einschätzung zu mittleren Reliabilitätskoeffizienten von SJTs zu geben und die Abhängigkeit der Messgenauigkeit von SJT- und Studiencharakteristika darzustellen. Im Rahmen des zweiten Artikels (Kapitel 2.2) steht die Konstruktvalidität von SJTs im Vordergrund. Es soll untersucht werden, inwieweit es möglich ist ein konstruktvalides Verfahren zu entwickeln, wenn man das klassische Verfahren zur Entwicklung und zum Scoring von SJT Items zugunsten eines strikt konstruktorientierten Ansatzes fallen lässt. Der dritte Artikel (Kapitel 2.3) untersucht schließlich die Fälschungsanfälligkeit dieses konstruktorientierten SJTs. Bisherige Studien zu Thema zeigen keine einheitliche Ergebnislage und weisen viele methodische Schwierigkeiten auf, die teilweise im erheblichen Ausmaß die Interpretierbarkeit der Ergebnisse einschränken. Im Rahmen von zwei Laboruntersuchungen wird daher das Ausmaß von Faking des SJTs gegenüber einem klassischen Selbstbeurteilungsfragebogen der Persönlichkeit analysiert.

Der dritte Teil der Synopsis schließt mit einer Gesamtdiskussion (Kapitel 3). Dabei werden sowohl die wichtigsten Erkenntnisse aus den Studien kurz zusammengefasst, als auch Limitationen diskutiert (Kapitel 3.1 bis 3.3). Des Weiteren soll eine Einordnung der Ergebnisse in die aktuellen praktischen und theoretischen Entwicklungen von SJTs gegeben werden (Kapitel 3.4). Die Synopsis schließt dementsprechend mit einem Ausblick auf potentielle Fragestellungen, die im Rahmen von Folgestudien erörtert werden können.

1.1 Definition und Struktur von SJTs

Wie bereits in Abbildung 1 verdeutlicht bestehen SJT Items aus der Beschreibung von berufsrelevanten und erfolgskritischen Situationen und einer Anzahl vorgegebener Handlungs-

alternativen, die der Bewerber anhand der spezifischen Instruktion evaluieren soll. Trotz der hohen Forschungsdichte finden sich in den Publikationen der letzten Jahrzehnte kaum etablierte Definitionen zu SJTs, die über diese bloße Beschreibungen des prototypischen Aufbaus eines SJT Items hinausgehen.

Konsens scheint aber zumindest darin zu bestehen, dass SJTs Messmethoden darstellen, die den simulationsorientierten Verfahren zuzurechnen sind. Im Allgemeinen dienen Simulationen im Rahmen von Auswahlprozessen dazu, Arbeitsgeschehen und -verhalten möglichst repräsentativ im Verfahren selbst abzubilden (Muck, 2013). Dementsprechend werden Bewerber schon während des Auswahlprozesses mit Situationen konfrontiert, die Schlüsselaspekte des Arbeitsplatzes abbilden (Lievens & De Soete, 2012; Tuzinski, 2013). Simulationsorientierte Verfahren unterscheiden sich untereinander dahingehend, inwieweit sie ein realitätsnahes Abbild des Arbeitsalltags ermöglichen. Dieses Ausmaß an Realitätsnähe wird auch als Fidelity bezeichnet. Zu den sogenannten high-fidelity Simulationen zählen z.B. Arbeitsproben und Rollenspiele, wie sie z.B. im Rahmen von Assessment Centern eingesetzt werden. Im Rahmen dieser Simulationen werden die Bewerber meist mit sehr realitätsnahen Szenarien und Aufgaben konfrontiert und das geforderte Verhalten orientiert sich stark an dem, was auch im tatsächlichen Berufsalltag gefordert ist. Im Vergleich zu diesen Simulationen, weisen SJTs deutlich weniger Fidelity auf (Kanning & Schuler, 2014). Motowidlo und Kollegen (Motowidlo et al., 1990; Motowidlo & Trippins, 1993) spezifizieren SJTs daher als *low-fidelity* Simulationen.

Ein weiterer Versuch der Spezifizierung findet sich z.B. bei McDaniel, Morgeson, Finnegan, Campion und Braverman (2001), die SJTs definieren als „*any paper-and-pencil test designed to measure judgment in work settings*“ (S. 730). Obwohl die Autoren diese Definition selbst als eher grobe Beschreibung bezeichnen, muss man festhalten, dass sie tatsächlich schon so spezifisch ist, dass sie der bestehenden Heterogenität in den Strukturmerkmalen von SJTs nicht gerecht wird. So sind SJTs nicht auf papierbasierte Versionen beschränkt. Es finden sich zunehmend auch multimediale SJTs, die Arbeitssituationen z.B. über kurze Videosequenzen darstellen. Zudem schwingt bei dieser Definition die Vorstellung mit, dass

SJTs testübergreifend spezifische Konstrukte messen, die sich hier z.B. unter dem Begriff der Urteilsfähigkeit (*judgment in work settings*) zusammenfassen lassen. Diese Sichtweise, dass SJTs Tests darstellen, die testübergreifend spezielle Fähigkeiten erfassen, die sich einem bestimmten Konstrukt zuordnen lassen, findet sich auch bei anderen Autoren. Zu den prominentesten Vertretern dieser Sichtweise zählen Sternberg und Kollegen (Hedlund & Sternberg, 2000; Stemler & Sternberg, 2006; Sternberg et al., 2000), die SJTs als Tests zur Messung der praktischen Intelligenz sehen. Allerdings stehen viele empirischen Ergebnisse gegen diese Sichtweise (vgl. McDaniel & Whetzel, 2005), so dass man sich heutzutage von dieser Vorstellung distanziert und SJTs als kontextualisierte Messmethoden definiert, die zur Erfassung verschiedenster Konstrukte entwickelt werden können (Lievens & Motowidlo, 2016).

1.2 Testcharakteristika

Wie gesehen gestaltet sich die Spezifikation einer Definition von SJTs sehr schwierig. Das hängt auch damit zusammen, dass SJTs eine sehr heterogene Gruppe von Messmethoden darstellen. Diese Heterogenität betrifft viele grundlegende Testcharakteristika von SJTs, angefangen bei der Entwicklung von Iteminhalten über das gewählte Antwortformat, bis hin zu Fragen der spezifischen Instruktion und der kognitiven Komplexität der Items. Hierbei handelt es sich nicht allein um phänotypische Kategorien, da die Wahl der spezifischen Testcharakteristika teilweise erhebliche Auswirkungen auf die psychometrischen Eigenschaften von SJTs haben. Im Folgenden werden daher einige Testcharakteristika kurz vorgestellt. Dabei handelt es sich um keine erschöpfende Darstellung, die aufgrund der weitreichenden Heterogenität auch schwer umsetzbar wäre, umfasst aber diejenigen Attribute, die als zentral angesehen werden (Champion et al., 2014). Dazu gehören die Entwicklung der Items, das Itemscoring, die Fidelity der Items, die Ausgestaltung des Antwortformats, die Komplexität der Items, und das Ausmaß des Kontextbezugs spezifischer SJTs.

Entwicklung. Bezogen auf die Entwicklung von SJT Items lassen sich vor allem zwei

grundlegende Richtungen unterscheiden: (a) der klassische, induktive und (b) der konstruktorientierte, deduktive Ansatz. Obwohl sich für letzteren in den letzten Jahren gesteigertes Interesse verzeichnen lässt, stellt der klassische Ansatz, gemessen an der Anwendungshäufigkeit, den weitaus prominenteren Ansatz dar. Er geht auf das klassische Konstruktionsprinzip nach Motowidlo et al. (1990) zurück. Hierbei besteht das primäre Ziel, Items zu entwickeln, die möglichst repräsentativ und erfolgskritisch für die vakante Stelle sind. Um dieses gewährleisten zu können, spielen Experten (sogenannte *subject matter experts*; SMEs) eine entscheidende Rolle. Sie werden sowohl zur Entwicklung realitätsnaher Itemstämme, als auch bei der Generierung realistischer Handlungsalternativen eingesetzt. Zur Entwicklung erfolgskritischer Situationen wird häufig die Methode der kritischen Ereignisse (Critical Incident Technique oder kurz CIT; Flanagan, 1954) als halbstandardisierte Methode der Anforderungsanalyse eingesetzt. Obwohl SJT Items, die nach diesem Ansatz entwickelt werden, ein (vermeintlich) inhaltsvalides Abbild der tatsächlich geforderten Arbeitsleistungen darstellen (Ployhart & MacKenzie, 2011), führt dieser Ansatz in der Regel auch zu konstruktheterogenen Verfahren. Das bedeutet, dass die einzelnen Items oder sogar die einzelnen Antwortoptionen ganz unterschiedliche Fähigkeiten und Fertigkeiten repräsentieren und auch in unterschiedlicher Weise mit Außenkriterien korrelieren. Wie ausgeprägt diese Multidimensionalität ist konnten McDaniel und Whetzel (2005; 2007) für die Antwortoptionen des in Abbildung 1 dargestellten Beispielitems zeigen. Hier wurden die Effektivitätseinschätzungen der Probanden mit Intelligenz, Gewissenhaftigkeit und Verträglichkeit korreliert. Für die einzelnen Antwortoptionen zeigen sich stark unterschiedliche Korrelationsmuster. Antwortoption 1 wird vor allem von solchen Probanden als besonders gut eingeschätzt, die hohe Intelligenzwerte (r = .10) und niedrige Verträglichkeit (r = -.13) aufweisen. Für Antwortoption 2 und 5 zeigen sich jeweils signifikante positive Korrelationen mit Intelligenz (r = .11 und r = .13). Antwortoption 3 wird hingegen vor allem von solchen Probanden hoch bewertet, die niedrige Verträglichkeitswerte aufweisen (r = -.16) und Antwortoption 4 weist eine negative Korrelation mit Intelligenz auf (r = -.17). Für Gewissenhaftigkeit zeigt keine der in Abbildung 1 dargestellten Antwortoptionen einen signifikanten Zusammenhang. Die Korrelationsmuster für weitere Items aus dem SJT fallen wiederum ganz anders aus und

erlauben deswegen wenig Einblick darin, welche Konstrukte mit dem SJT erfasst werden.

Um diesen Mangel an Konstruktspezifikation des klassischen Entwicklungsparadigmas nach Motowidlo et al. (1990) zu kompensieren, zeichnet sich der deduktive Ansatz vor allem dadurch aus, dass die Items auf *a priori* definierte Konstrukte hin entwickelt werden, die entweder aus der Theorie abgeleitet sind oder das Ergebnis einer gründlichen Arbeitsanalyse darstellen (Weekley, Ployhart & Holtz, 2006). Obwohl die Messung spezifischer Konstrukte mit vielen Vorteilen verbunden ist, findet der deduktive Ansatz noch immer vergleichsweise wenig Anwendung (Christian, Edwards & Bradley, 2010; Ployhart & MacKenzie, 2011). Das mag auch damit zusammenhängen, dass für diesen Ansatz noch keine allgemeinen Handlungsempfehlungen definiert sind, wie man genau bei der Entwicklung vorgehen sollte. Zudem fehlen zumeist anschließende Untersuchungen zum Nachweis von Konstruktvalidität, was eine Beurteilung der Brauchbarkeit der verschiedenen Vorgehensweisen erschwert, bzw. ausschließt.

Scoring. Hat man SJT Items entwickelt, so stellt sich die Frage nach dem Scoring, d.h. wie man die Antworten der Bewerber bewertet. Diese Frage ist bei SJTs von besonderer Bedeutung, da hier die Items in der Regel keine expliziten, objektiv korrekten Antwortoptionen aufweisen (Bergman, Drasgow, Donovan, Henning & Juraska, 2006). Die in Abbildung 1 dargestellten Handlungsalternativen verdeutlichen die Schwierigkeit, die mit der Bewertung von Probandenantworten einhergehen, da von den verschiedenen Optionen keine auf den ersten Blick komplett unplausibel erscheint. Daher ist es von entscheidender Bedeutung die einzelnen Antwortoptionen hinsichtlich ihrer Effektivität zu bewerten. Die Quelle dieser Bewertung unterscheidet sich dabei in den verschiedenen Scoringmethoden, von denen in der Literatur vorrangig drei verschiedene Vorgehensweisen unterschieden werden: (1) der experten-basierte Ansatz, (2) der empirische Ansatz und (3) der theoretische, konstruktorientierte Ansatz. Auch Kombinationen dieser Ansätze in Form von hybriden Scoringschlüssel sind denkbar. Der am häufigsten angewandte Ansatz stellt der experten-basierte Ansatz nach Motowidlo et al. (1990) dar, bei dem wiederum SMEs eingesetzt werden, um die Effektivität der einzelnen Antwortalternativen einzuschätzen. Die Antworten der Bewerber werden dann

dahingehenden bewertet, inwieweit sie mit der Meinung der SMEs übereinstimmen.

Der empirische Ansatz leitet sich aus Scoringverfahren ab, die vor allem im Rahmen der Auswertung von biografischen Verfahren bekannt sind. Hierbei wird die Effektivität der einzelnen Antwortoptionen anhand ihres Zusammenhangs mit einem Außenkriterium bestimmt (Mount, Witt & Barrick, 2000; Mumford, 1999). Antwortoptionen werden demnach höher bewertet, wenn diese gut zwischen geeigneten und ungeeigneten Bewerbern differenzieren können. Allerdings ist dieses Vorgehen stark atheoretisch und vor allem deswegen auch immer wieder kritisch diskutiert, z.B. bezüglich der fragwürdigen Stabilität und Generalisierbarkeit des resultierenden Scoring-Schlüssels (Bergman et al., 2006).

Im Rahmen des theorie- bzw. konstruktorientierten Vorgehens besteht schließlich das primäre Ziel darin, konstrukthomogene SJT Items zu generieren. Dementsprechend werden die einzelnen Antwortoptionen dahingehend bewertet, inwieweit sie Verhalten repräsentieren, das mit einem bestimmten Trait assoziiert ist, also z.B. in welchem Ausmaß die einzelnen Antwortoptionen verträgliches Verhalten repräsentieren. Über dieses Vorgehen kann ein generelles Verständnis davon entwickelt werden, warum bzw. in welchen Situationen ein SJT berufsrelevante Leistungskriterien vorhersagen kann. Allerdings wird auch vermutet, dass konstruktorientierte Scoring-Schlüssel aufgrund erhöhter Transparenz ein höheres Risiko an Verfälschbarkeit aufweisen (Bergman et al., 2006; Kluger, Reilly & Russel, 1991)

Fidelity. Wie bereits erwähnt, werden SJTs den simulationsorientierten Verfahren zugerechnet, wählen aber als sogenannte *low-fidelity Simulationen* einen weniger realitätsnahen Ansatz der Simulation als z.B. Arbeitsproben oder Rollenspiele. Obwohl sich SJTs also schon per Definition am unteren Ende des Fidelity-Kontinuums verortet lassen (Munshi, Lababidi & Alyousef, 2015), zeigen sich zwischen verschiedenen SJTs trotzdem noch Unterschiede bezogen auf die Realitätsnähe der Items. Diese Unterschiede beschränken sich auf die Fidelity der Stimuluskomponente von SJTs, da die Antwortalternativen fast ausschließlich schriftlich

präsentiert werden[2]. Das Ausmaß der Fidelity eines spezifischen SJTs ist in hohem Maße abhängig von dem Medium, über das die Situationsbeschreibung den Bewerbern präsentiert wird. Neben den klassischen Papier-Bleistift Formaten gibt es auch zunehmend multimediale Präsentationsmöglichkeiten. So können problematische Situationen z.B. über kurze Videosequenzen dargestellt werden, die es dem Anwender ermöglichen sehr viel direkter die Interaktion zwischen den beteiligten Personen wahrzunehmen, als das in der papierbasierten Variante möglich ist. Obwohl diese Präsentationsform häufig stärker an einen spezifischen Anwendungskontext gebunden sind und mit einem höheren Aufwand bezogen auf monetäre und administrative Ressourcen einhergeht, werden sie von den Bewerbern zumeist als stärker berufsbezogen wahrgenommen (Richman-Hirsch, Olson-Buchanan & Drasgow, 2000), was sich positiv auf die wahrgenommene Fairness und Akzeptanz seitens der Bewerber auswirken sollte (Bauer & Truxillo, 2006).

Antwortformat. Bei der Gestaltung des Antwortformates geht es um die konkrete Ausgestaltung der Aufgabenstellung (Muck, 2013). Zwischen verschiedenen SJTs zeigen sich dabei sowohl Unterschiede bezogen auf die Antwortinstruktionen als auch bezüglich des Bewertungsformats, die im folgenden kurz dargestellt werden sollen. Bezüglich der verschiedenen Instruktionen wird häufig eine dichotomisierte Taxonomie angeführt nach der sich *wissensbasierte* und *verhaltensbasierte* Instruktionen voneinander unterscheiden lassen (Ployhart & Ehrhart, 2003). Bei ersterer steht die Frage nach der Effektivität oder Qualität der Handlungsoptionen im Vordergrund. Sie werden dementsprechend auch unter der Bezeichnung „should do"-Instruktionen gefasst und sind geeignet maximales Verhalten zu erfassen. Die in Abbildung 1 verwendete Antwortinstruktion folgt genau diesem Muster. Im Vergleich dazu fragen verhaltenbasierte Instruktionen nach persönlichen Präferenzen und

[2]Ausnahmen bilden hierbei einige wenige SJTs, die die Handlungsalternativen über Videosequenzen präsentieren (Lievens & De Soete, 2015) und sogenannte Webcam Tests (Oostrom, Born, Serlie & van der Molen, 2011), die sich durch ein offenes Antwortformat auszeichnen. Da beide Ansätze allerdings nur sehr selten zum Einsatz kommen und bei Webcam Tests außerdem strittig ist, ob sie den SJTs überhaupt zugerechnet werden können oder nicht doch eher ein eigenständiges Messverfahren darstellen, werden diese im Folgenden nicht weiter behandelt

zielen damit auf typisches Verhalten. Sie werden kurz auch als „would do"-Instruktionen bezeichnet. Die Wahl der Instruktion hat dabei entscheidende Konsequenzen auf die Konstruktvalidität. Wie die Unterscheidung in maximales und typisches Verhalten vermuten lässt, konnten McDaniel, Hartmann, Whetzel, und Grubb (2007) metaanalytisch zeigen, dass wissensbasierte SJTs einen höheren Zusammenhang mit kognitiven Fähigkeiten aufweisen als SJTs, die verhaltensbasierte Instruktionen verwenden (ρ=.35 für wissenbasierte und ρ=.19 für verhaltensbasierte SJTs), während letztere dementsprechend höher mit den Persönlichkeitstraits Verträglichkeit (ρ=.19 für wissenbasierte und ρ=.37 für verhaltensbasierte SJTs), Gewissenhaftigkeit (ρ=.24 für wissenbasierte und ρ=.34 für verhaltensbasierte SJTs), und Emotionale Stabilität (ρ=.12 für wissensbasierte und ρ=.35 für verhaltensbasierte SJTs) korrelieren.

Bei der Frage nach dem Bewertungsformat geht es darum, in welcher Form der Proband die Antwortoption bewerten soll. Dazu lassen sich in der Literatur vor allem Forced-Choice- und Ratingformate unterscheiden. Andere Verfahren, wie z.B. Rankingformate werden so selten angewendet, dass sie kaum eine Rolle spielen. In dem dargestellten Beispielitem (Abbildung 1) wurde ein Ratingformat verwendet. Dementsprechend sind die Probanden aufgefordert jede der dargestellten Antwortoptionen anhand einer siebenstufigen Skala zu bewerten. Demgegenüber würde bei einem Forced-Choice Format keine vollständige Bewertung jeder Antwortoption stattfinden. Es findet vielmehr eine vergleichende Bewertung der Handlungsalternativen statt. Der Proband wäre dementsprechend z.B. aufgefordert, aus den dargestellten Handlungsoptionen diejenige auszuwählen, die seiner Meinung nach die effektivste in der gegebenen Situation darstellt, bzw. diejenige, die seinem individuellen Verhalten am wahrscheinlichsten entspricht. Dieses sogenannte ‚*Pick Best*'-Verfahren kann auch noch erweitert werden, indem der Proband zusätzlich instruiert wird auch die Option anzugeben, die seiner Meinung nach die ineffektivste, bzw. die für ihn unwahrscheinlichste darstellt (Weekley et al., 2006).

Kognitive Komplexität: SJT Items weisen teilweise erhebliche Unterschiede in ihrer kognitiven Komplexität auf, d.h. dem Ausmaß in dem die Bearbeitung intellektuelle

Fähigkeiten von den Probanden abverlangen. Dabei scheint vor allem die Komplexität der Situationsbeschreibung von Bedeutung, während sich aus der bestehenden Literatur keine Konsequenzen aus der Komplexität der Handlungsoptionen ableiten lassen (Weekley et al., 2006). Kognitive Komplexität stellt kein eindimensionales Konstrukt dar, sondern setzt sich vielmehr aus verschiedenen Faktoren zusammen. Faktoren, die im Zusammenhang mit der Komplexität der Itemstämme genannt werden, sind die Länge der Situationsbeschreibung (*stem length*), die Komplexität der dargebotenen Interaktionen (*stem complexity*) und die Verständlichkeit der Situationsbeschreibung (*stem comprehensibility*; vgl. McDaniel & Whetzel, 2007; McDaniel, Whetzel & Nguyen, 2006). Diese Faktoren lassen sich nur schwerlich unabhängig voneinander beurteilen, da sie häufig zusammenhängen. Als Beispiel für ein hochkomplexes SJT kann z.B. das *Tacit Knowledge Inventory for Managers* (TKIM, Wagner & Sternberg, 1991) angeführt werden, bei dem die Situationsbeschreibungen typischerweise mehrere Absätze einnehmen und sehr komplexe Interaktionen beinhalten (McDaniel, Whetzel, Hartman, Nguyen & Grubb, 2006). Da der Einfluss der verschiedenen Faktoren der kognitiven Komplexität u.a. auf die erhöhten Anforderungen auf das Leseverständnis zurückgeführt werden, haben diese häufig einen nicht unerheblichen Einfluss auf Subgruppendifferenzen, spielen allerdings nur bei schriftlich dargebotenen Situationsbeschreibungen eine Rolle (Lievens & Sackett, 2006).

Kontextbezug. Der Kontextbezug ist für SJT Items definitorisch, da die verschiedenen Handlungsalternativen vor dem Hintergrund der dargebotenen Situation bewertet werden sollen (Lievens, 2006; Lievens & De Soete, 2012, 2015). Allerdings unterscheidet sich das Ausmaß dieses Kontextbezugs zwischen verschiedenen SJTs in dem Grad, indem ein SJT an die Spezifika einer einzelnen Stelle gebunden sind. Die Frage nach dem Kontextbezug hat dementsprechend vor allem Relevanz für die generalisierte Anwendbarkeit einzelner Verfahren. So werden einige SJTs im Rahmen spezifischer Auswahlverfahren entwickelt und zeigen daher einen entsprechend hohen Bezug zu den Eigenheiten der vakanten Stelle. Eine Anwendung in anderen Kontexten ist dementsprechend nur sehr stark eingeschränkt möglich, bzw. sogar vollständig ausgeschlossen. Wiederum andere SJTs, vor allem

solche, die eher konstrukthomogene Verfahren darstellen, beziehen überwiegend generische Situationsbeschreibungen mit ein, um die Anwendung des Verfahrens in vielen verschiedenen Kontexten gewährleisten zu können. Beispiele für solche Testverfahren sind z.B. der *Mayer-Salovey-Caruso Test zur Emotionalen Intelligenz* (MSCEIT, in der deutschen Adaption von Steinmayr, Schütz, Hertel & Schröder-Abé, 2011), der *Situational Test of Emotion Managment* (STEM, MacCann & Roberts, 2008), oder der Situationsbeurteilungsfragebogen im *Test zur Messung interkultureller Kompetenz* (TMIK; Schnabel, Kelava, Seifert & Kuhlbrodt, 2014). Zwischen diesen Extremen gibt es natürlich noch eine Reihe von Abstufungen, die z.B. solche SJTs umfassen, die für eine Gruppe von Stellen geeignet ist. Das in Abbildung 1 dagestellt Item gehört z.B. zu einem SJT, dass zur Auswahl von Managern mit Weisungsbefugnis in vielen Kontexten angewendet werden kann.

1.3 Psychometrische Eigenschaften von SJTs

Wie bereits in der Einleitung geschildert, zeigt sich seit den 1990er Jahren ein stark gesteigertes Interesse an der Anwendung von SJTs in verschiedensten Anwendungskontexten. Dieses lässt sich vor allem damit begründen, dass in zahlreichen empirischen Untersuchungen und Metaanalysen eine Reihe positiver Eigenschaften von SJTs nachgewiesen werden konnten. Wenig verwunderlich für eine Methode, die vorrangig im Rahmen der Personalauswahl eingesetzt wird, liegt hierbei der primäre Forschungsschwerpunkt - gemessen an der Anzahl der Publikationen - im Nachweis der Vorhersageleistung von leistungsrelevanten Kriterien durch SJTs. In Metaanalysen zu diesem Thema zeigten sich große bis mittlere korrigierte Validitätkoeffizienten von ρ=.34 (McDaniel et al., 2001) bzw. ρ=.26 (McDaniel et al., 2007), die vergleichbar sind mit den prädiktiven Leistungen von anderen, weitaus etablierteren Verfahren der Personalauswahl (z.B. Assessment Centern; vgl. Schmidt & Hunter, 1998). Zudem lassen sich auf Basis von SJTs sogar noch nach neun Jahren valide Unterschiede in berufsrelevanten Leistungskriterien vorhersagen (Lievens & Sackett, 2012). In der Metaanalyse von McDaniel et al. (2007) konnte zudem der inkrementelle Nutzen von SJTs im Rahmen

von Auswahlverfahren dargestellt werden: Gegenüber kognitiven Fähigkeiten klären sie 3 bis 5 % zusätzliche Varianz auf, 6 bis 7 % im Vergleich zu Persönlichkeit und 1 bis 2 % im Vergleich zu einer Kombination beider Konstrukte. Obwohl diese Koeffizienten kleiner sind, als solche, die man mit Hilfe von sogenannten *high-fidelity* Simulationen (z.B. Assessment Centern) findet, kann man festhalten, dass Unternehmen mit SJTs das Spektrum der erfassten eignungsrelevanten Merkmale (*knowledge, skills, abilities and other characteristics; KSAOs*) über traditionelle Persönlichkeits- und Fähigkeitsmaße hinaus erweitern können. Darüber hinaus erzeugen SJTs weniger Subgruppendifferenzen als kognitive Leistungstests (Whetzel & McDaniel, 2009) und haben durch den Simulationsansatz eine wahrgenommene berufliche Relevanz, die sich positiv auf die Bewerberreaktion und -akzeptanz auswirkt (Kanning, Grewe, Hollenberg & Hadouch, 2006). Im direkten Vergleich zu anderen simulationsorientierten Verfahren, wie Arbeitsproben oder Assessment-Centern, weisen sie zudem entscheidende ökonomische Vorteile auf, bezogen sowohl auf die monetären als auch auf die zeitlichen Ressourcen, die vor allem mit der Administration und Auswertung eines solchen Verfahrens einhergehen (Weekley & Ployhart, 2006).

Allerdings weisen SJT erhebliche Einschränkungen bezogen auf den Nachweis anderer grundlegegender psychometrischer Eigenschaften auf. So lassen sich mit SJT nur selten Reliabilitäten finden, die den gängigen Daumenregeln zur Anwendung in Auswahlkontexten gerecht werden (vgl. Nunnally & Bernstein, 1994), und Validitätsüberprüfungen beschränken sich zum Großteil auf die Analyse der prognostischen Vorhersagekraft von SJTs. Fragestellungen, die auf die Konstruktvalidität von SJTs zielen, finden nur wenig Beachtung im Rahmen von empirischen Untersuchungen. Sowohl die Nützlichkeit im Rahmen von praktischen Anwendungen, als auch die wissenschaftliche Bewertung von SJTs ist dementsprechend stark eingeschränkt. So macht es der oftmals fehlende Konstruktbezug auch schwierig, SJTs mit anderen Verfahren der Eignungsdiagnostik zu vergleichen, z.B. bezüglich der Verfälschbarkeit. Da genau diese Unzulänglichkeiten - sprich die Messgenauigkeit, die Konstruktvalidität und die Fälschungsanfälligkeit von SJTs - Ausgangspunkte der empirischen Untersuchungen darstellen, werden diese im folgenden ausführlicher dargestellt werden.

1.3.1 Reliabilität von SJTs

Der Nachweis angemessener Messgenauigkeit eines Verfahrens ist eine *conditio sine qua non* des psychologischen Testens (Vacha-Haase & Thompson, 2011). Vor allem bei Verfahren mit denen Individualdiagnosen und -entscheidungen getroffen werden sollen, werden erhebliche Anforderungen an die Messgenauigkeit gestellt (Nunnally & Bernstein, 1994). Obwohl SJTs verstärkt im Rahmen von Auswahlverfahren eingesetzt werden und auch in Forschungskontexten häufig Anwendung finden, gibt es bisher noch unzureichende Erkenntnisse dazu, inwieweit es möglich ist mit SJTs reliable Ergebnisse zu erzeugen. Dabei zeigen sich viele problematische Aspekte in der bestehenden Literatur. Erstens werden in vielen Studien gar keine Reliabilitätsschätzungen zu dem verwendeten SJT angegeben. In der Studie von McDaniel et al. (2001) wird das Ausmaß dieses Problems verdeutlicht: Aus den 72 Studien, die sie im Rahmen ihrer Validitätsgeneralisierung analysiert haben, konnten sie gerade einmal 33 Reliabilitätskoeffizienten extrahieren. Nicht einmal die Hälfte der Studien machte dementsprechend Angaben zur Messgenauigkeit des verwendeten SJTs. Bedenkt man hierbei, dass der Messfehler die Schätzung von Effektstärken beeinflusst, und in nicht unerheblichen Maße Einfluss auf viele statistische Verfahren nimmt (Schmidt, Hunter & Urry, 1976; Schmitt & Chan, 2006), sollte die Berechnung der Reliabilität am Anfang jeder quantitativen Studie stehen (Vacha-Haase & Thompson, 2011; Wilkinson & American Psychological Association Task Force on Statistical Inference, 1999).

Darüber hinaus vermitteln die wenigen Studien, in denen Schätzungen der Reliabilität berichtet werden, ein problematisches Bild: Nur wenige Reliabilitätskoeffizienten belegen ein ausreichendes Ausmaß an Messgenauigkeit und Catano, Brochu & Lamerson (2012) fanden in einer Analyse von 56 internen Konsistenzschätzungen nur ein mittleres α von .46. Allerdings zeigen sich in empirischen Studien häufig eine große Variabilität bezogen auf die angegebenen Reliabilitätskoeffizienten. Da Catano et al. (2012) selbst keine Angaben dazu machen, lässt sich nicht beurteilen, ob der von ihnen angegebene mittlere Wert eine homogene Schätzung darstellt. Bedenkt man jedoch die vielen Charakteristika, in denen SJTs sich unterscheiden,

so scheint diese Annahme zumindest fragwürdig. In der bestehenden empirischen Literatur lassen sich allerdings keine Ergebnisse bezüglich des Einflusses von Testcharakteristika auf die Messgenauigkeit finden, die über den Einfluss der Itemanzahl hinausgehen. Hier zeigt sich der zu erwartende Effekt, dass die Itemanzahl in einem positiven Verhältnis zu der Messgenauigkeit eines SJTs steht (Catano et al., 2012). Bedenkt man allerdings die vielen offenen Fragen, die sich zur Messgenauigkeit von SJTs stellen, so scheint eine differenziertere Aussage sowohl zur gängigen Dokumentationspraxis von Reliabilitätskoeffizienten als auch zu typischen Reliabilitäten und deren Abhängigkeit von essentiellen Testcharakteristika dringend notwendig.

1.3.2 Konstruktvalidität von SJTs

Da SJTs vorrangig im Rahmen von Auswahlprozessen angewendet werden, in denen das primäre Ziel zumeist in der Vorhersage von gegenwärtiger und/oder zukünftiger Arbeitsleistung liegt, tritt der Konstruktaspekt von SJTs häufig in den Hintergrund. Dementsprechend gibt es im Vergleich zu der umfassenden Literatur, die sich mit der Kriteriumsvalidität von SJTs auseinandersetzt, nur wenige Untersuchungen, die sich mit der Konstruktvalidität beschäftigen (McDaniel, List & Kepes, 2016; Sorrel et al., 2016). In ihrer Metaanalyse stellen Christian et al. (2010) das Ausmaß des Problems dar: Ein Drittel (45 von 136) der publizierten Studien, die in ihre Analyse eingingen, gaben keinerlei Angaben zu den erfassten Konstrukten. Darüber hinaus muss man festhalten, dass bei denjenigen Artikeln, bei denen explizit auf Konstrukte verwiesen wird, zumeist die entsprechende empirische Evidenz fehlt, das diese auch tatsächlich erfasst werden (Ployhart & MacKenzie, 2011). Zusammengenommen bedeutet dies, dass SJTs in vielen Kontexten angewendet werden, ohne das genauer spezifiziert wird, welche Konstrukte diese Verfahren in dem spezifischen Fall messen. Dieses „methodenorientierte“ Vorgehen findet sich zwar auch bei anderen Verfahren, die verstärkt in der Personalauswahl eingesetzt werden (wie z.B. Arbeitsproben oder Interviews), allerdings greift allein das Wissen, dass mit Hilfe eines Verfahren in einem Auswahlkontext valide Per-

sonalentscheidungen getroffen werden können, zu kurz. Auch das „wie“ und das „warum“ sind entscheidende Fragestellungen in diesem Kontext, die aber nur über die Identifikation der zugrundeliegenden Konstrukte hinreichend aufgeklärt werden können. Dabei ist der Nachweis von Konstruktvalidität nicht einzig ein epistomologisches Bestreben, sondern ist mit vielen Vorteilen assoziiert (Christian et al., 2010; Lievens, Buyse & Sackett, 2005; Sorrel et al., 2016; Trippe, 2002). So konnten Christian et al. (2010) zeigen, dass sich ein grundlegendes Verständnis der erfassten Konstrukte über die Möglichkeit einer theoriegeleiteten Kriteriumswahl positiv auf die Vorhersageleistung von SJTs auswirken kann. Zudem kann ein generelles Verständnis davon, was gemesssen wird auch den Anwendungskontext von SJTs erweitern. So könnte ein konstruktvalides Verfahren auch als Feedbackinstrument im Rahmen der Personalentwicklung oder bei Trainingsmaßnahmen eingesetzt werden.

Bisherige empirische Untersuchungen zur Konstruktvalidität von SJT, die z.B. die interne Struktur von SJTs über explorative Faktorenanalysen überprüfen, führen fast immer zu der Extraktion einer großen Zahl an Faktoren, die sich häufig nur schwer interpretieren lassen (Lievens, Peeters & Schollaert, 2008; McDaniel et al., 2016; Sorrel et al., 2016). Ein großes Problem, das hier zum Ausdruck kommt und das dem Nachweis von Konstruktvalidität entgegensteht, ist die inhärente Multidimensionalität auf Item-Ebene, die einem SJT in der Regel attestiert wird. Bemühungen, den Konstruktaspekt genauer zu verstehen, beschränken sich daher in der Regel auf korrelative Untersuchungen, um das Verhältnis der Ergebnisse bestehender SJTs zu anderen Konstrukten - vor allem kognitive Fähigkeiten und Persönlichkeit - zu untersuchen. Obwohl dieses Vorgehen im Rahmen von Validierungsstudien bei der Bestimmung von konvergenter und diskriminanter Validität nicht ungewöhnlich ist, ist dieses Vorgehen bei SJTs häufig nicht an die theoretischen Vorstellungen im Sinne eines nomologischen Netzes geknüpft, d.h. es erfolgt zumeist atheoretisch und ohne die Berücksichtigung der Konstrukte, die man mit dem spezifischen SJT messen möchte. Dementsprechend begrenzt sind die Erkenntnisse, die aus solchen Untersuchungen gezogen werden können, vor allem da in solchen Untersuchungen zuweilen auch sehr kontraintuitive Ergebnisse herauskommen, die nur wenige Rückschlüsse auf die tatsächlich erfassten Konstrukte erlauben.

Wenn man SJTs vor dem Hintergrund eines konstruktorientierten Ansatzes verstehen und bewerten möchte, kann der Weg nicht darin bestehen ein konstrukttheterogenes SJT zu nehmen und über Korrelationen zu versuchen *a posteriori* Dimensionen zu identifizieren. Auf Grundlage der empirischen Studien und der bestehenden Literatur lässt sich zumindest festhalten, dass sich der klassische Entwicklungs- und Scoringansatz nach Motowidlo et al. (1990) aus mehreren Gründen eher wenig dazu eignet, konstrukthomogene Items hervorzubringen (Weekley et al., 2006). Erstens sind die Iteminhalte im Regelfall über die Methode der kritischen Ereignisse (Flanagan, 1954) abgeleitet. Primäres Ziel besteht hierbei über die Rekonstruktion erfolgskritischer Situationen und angelehnter Handlungsweisen die Entwicklung eines inhaltsvaliden Verfahrens. Dementsprechend werden bei der Entwicklung keine zuvor definierten Konstrukte berücksichtigt. Darüber hinaus können sich die daraus resultierenden Situationsbeschreibungen stark in ihrer kognitiven Komplexität unterscheiden und dementsprechend im Rahmen der Anwendung mehr oder weniger anfällig gegenüber konstruktirrelevanter Varianz sein. Zweitens sind die Antwortoptionen, die innerhalb einer Situationsbeschreibung genestet sind, nicht konstrukthomogen. Mit Hilfe von Experten werden Antwortalternativen zu einer Situation genereriert, die häufig eine Vielzahl verschiedenster Handlungen repräsentieren, die in unterschiedlicher Weise mit Konstrukten korrelieren. Diese Multidimsionalität auf Itemebene wurde ja bereits für das Beispielitem (Abbildung 1) dargelegt. Drittens werden für die Entwicklung eines Bewertungsschlüssels Experten eingesetzt, denen bei der Evaluation der Effektivität der einzelnen Handlungsoptionen keine spezifischen Konstrukte an die Hand gegeben werden. Dadurch ist nicht auszuschließen, dass sich sowohl eine starke interindividuelle Heterogenität zwischen den Experten, als auch eine intraindividuelle Heterogenität für jedes Item oder sogar jede Antwortoption bezüglich der Bewertungskriterien zeigt. Dementsprechend legen die empirischen und theoretischen Erkenntnisse den Schluss nahe, dass man zur Entwicklung eines konstruktvaliden SJTs nicht um die Anpassung des klassischen Entwicklung- und Scoringansatzes umhin kommt.

1.3.3 Verfälschbarkeit von SJTs

Wie bei anderen Verfahren, die im Rahmen von Auswahlprozessen eingesetzt werden, stellt sich auch bei SJTs die Frage, inwieweit Bewerber ihre Ergebnisse durch Antworttendenzen im Sinne von Faking systematisch verzerren, d.h. inwieweit sich Personen geeigneter darstellen können, als sie tatsächlich sind. Aus organisatorischer Sicht, aber auch aus individueller Bewerberperspektive, lassen sich eine Reihe von Gründen unterscheiden, warum die Verfälschbarkeit von Testergebnissen potentiell problematisch ist. Erstens stellt sich die Frage, inwieweit sich Faking negativ auf die Kriteriumsvalidität eines Verfahrens auswirkt (Viswesvaran & Ones, 1999) und zweitens besteht das Risiko, dass die Rangordnung der Bewerber durch Faking verändert wird (Sackett, 2012). In diesem Fall hätte die Verfälschung einen erheblichen Einfluss auf individuelle Auswahlentscheidungen. Die empirischen Untersuchungen, die sich mit der Fälschungsanfälligkeit von SJTs befassen, liefern stark heterogene Ergebnisse, die eine generalisierte Betrachtungsweise erschweren. Hooper, Cullen und Sackett (2006) stellen diese Variabilität der Ergebnisse in Fakingsstudien in einem Review zusammen. So liegen in Laborstudien die standardisierten Mittelwertsunterschiede zwischen Personen, die instruiert sind ehrlich zu antworten und solchen die sich als besonders geeignet darstellen sollen, zwischen $d = 0.08$ und $d = 0.89$. Im Rahmen von Felduntersuchungen fällt die beobachtete Variabilität sogar noch höher aus: Vergleicht man die Antworten tatsächlicher Bewerber mit denen von Stelleninhabern so liegen die standardisierten Mittelwertsdifferenzen zwischen $d =$ - 0.60 und $d = 0.88$.

Obwohl es aufgrund dieser Variabilität in den Studienergebnissen schwer fällt, das generelle Ausmaß an Faking im Zusammenhang mit SJTs zu spezifizieren, gibt es Hinweise darauf, dass SJTs weniger fakinganfällig sind als klassische Persönlichkeitstests (Hooper et al., 2006; Kanning & Kuhne, 2006; Nguyen, Biderman & McDaniel, 2005). So fanden Nguyen und Kollegen (2005) in ihrer Studie mittlere bis große Faking-Effekte für das verwendete Persönlichkeitsinventar (Effektstärken lagen im mittel zwischen $d = 0.36$ für Verträglichkeit und $d = 0.76$ für Emotionale Stabilität), aber nur kleine Effekte für das verwendete SJTs

(mittleres $d = 0.15$). Ein fundamentales methodisches Problem, dass das Verständnis und den Nutzen dieser Studien einschränkt, ist allerdings die fehlende die Möglichkeit die Effekte der Messmethode von denen der gemessenen Konstrukte zu trennen. Da die verwendeten SJT konstruktheterogene Verfahren darstellen, die nach dem klassischen „methodenorientierten" Verfahren entwickelt worden sind, kann man demnach nicht beurteilen, ob Unterschiede in der Verfälschbarkeit auf die unterschiedlichen Methoden oder auf Unterschiede in den erfassten Konstrukten zurückzuführen sind.

Zusammenfassend lässt sich also festhalten, dass ein verbessertes Verständnis bezüglich der Fälschungsanfälligkeit von SJT nur über ein verbessertes konzeptuelles Verständnis von SJTs führen kann. Kann man spezifizieren, welche Konstrukte gemessen werden, so lassen sich Unterschiede in der Fakinganfälligkeit zwischen verschiedenen SJTs aber auch im Vergleich zu anderen Messverfahren interpretieren. Empirische Studien zu solchen Vergleichen finden sich allerdings bisher noch nicht.

2 Die vorliegenden Arbeiten

Die vorliegenden empirischen Arbeiten setzen vor allen an den Punkten an, die in der bisherigen SJT Literatur kaum oder keine Beachtung gefunden haben. Grundlage bilden daher diejenigen psychometrischen Eigenschaften, auf die in den vorangegangenen Abschnitten gesondert eingegangen worden ist. Im Rahmen der ersten empirischen Arbeit sollen die Möglichkeiten untersucht werden, mit Hilfe von SJTs reliable Ergebnisse zu erzielen. Des Weiteren wird der Einfluss grundlegender Test- und Studiencharakteristika auf die internen Konsistenzschätzungen untersucht. Im Rahmen des zweiten Artikels steht die Konstruktvalidität von SJTs auf dem Prüfstand. Es wird untersucht, inwieweit sich durch die Anpassung des Entwicklung- und Scoringverfahrens ein konstruktvalides Verfahren zur Messung von Persönlichkeitskonstrukten erzeugen lässt. In Artikel 3 liegt der Fokus auf der Fälschungsanfälligkeit dieses konstruktorientierte SJTs. Im Rahmen von zwei Studien soll sowohl das generelle Ausmaß der Verfälschbarkeit quantifiziert werden, als auch in Vergleich gesetzt werden zu der Fälschungsanfälligkeit eines klassischen Selbstbeurteilungsfragebogen der Persönlichkeit. Potentielle Unterschiede zwischen den beiden Verfahren sollen zudem unter Rückgriff auf die bestehende Fakingliteratur untersucht und erklärt werden.

2.1 Artikel 1: SJTs und Reliabilität

Originalpublikation: Kasten, N. & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of situational judgment tests (SJTs). *European Journal of Psychological Assessment, 32,* 230-240.

2.1.1 Einleitende Bemerkung

Wie bereits im theoretischen Teil der Arbeit dargestellt, bestehen einige Probleme, wenn es darum geht, das Ausmaß des Messfehlers bei SJTs systematisch zu quantifizieren. Ziele der vorliegenden Studie sind entsprechend der zugrundeliegenden Problematik mannigfaltig. Erstens soll dargestellt werden, in welchem Ausmaß Reliabilitätsschätzungen zu SJTs innerhalb der empirischen Untersuchungen überhaupt berichtet werden. Zweitens soll spezifiziert werden, welche Form der Reliabilitätsschätzung genau berichtet wird, da sich mit Hilfe der verschiedenen Schätzverfahren auch unterschiedliche Quellen von Messfehlern analysieren lassen (Nimon, Zientek & Henson, 2012). Drittens sollen zudem die mittlere Reliabilitätsschätzung als auch die Variabilität der beobachteten Koeffizienten quantifiziert werden. Bei signifikanter Variabilität sollen zudem Merkmale der Studien und des verwendeten SJTs als Moderatorvariablen einbezogen werden. Dabei sollen möglichst alle der dargestellten Testcharakteristika berücksichtigt werden, um so ein differenziertes Bild der Messgenauigkeit von SJTs zu bekommen.

Als Auswertungsmethode wurde die metaanalytische Methode der Reliabilitätsgeneralisierung (Vacha-Haase, 1998) gewählt. Um der genesteten Struktur der Daten gerecht zu werden (d.h. Koeffizienten sind genestet in Studien), wurde ein hierarchisch lineares Modell (HLM; Raudenbush & Bryk, 2002) mit drei Ebenen verwendet.

2.1.2 Abstract

During the past 20 years, Situational Judgment Tests (SJTs) have developed into a viable tool in personnel selection. Despite their growing popularity, research examining the extent of measurement error is widely lacking. Using reliability generalization, the aim of this article was twofold: (1) establish an estimate for an average coefficient alpha of SJT scores across studies and (2) examine the influence of essential SJT features and selected study variables on score reliability. To handle potential dependent observations a three-level hierarchical linear model was used. The results indicate that the reliability of SJT scores is typically rather low and below recommended levels for high-stakes applications. Additionally, both SJT and study characteristics affect score reliability. Implications for practitioners and researchers are provided to guide an appropriate use of SJTs and to initiate future research.

2.2 Artikel 2: SJTs und Konstruktvalidität

Originalpublikation: Kasten, N. (eingereicht). Development and validation of a situational judgment test (SJT) to measure Big Five personality dimensions.

2.2.1 Einleitende Bemerkung

Neben der angemessenen Messgenauigkeit stellt vor allem der oftmals fehlende Nachweis der Konstruktvalidität eine Krux in der Anwendung von SJTs dar. In der Einleitung wurde bereits hinreichend dargelegt, wie groß dieses Problem bei SJTs ist, und welche Konsequenzen daraus resultieren. Darüber hinaus wurde auch dargelegt, dass das bisherige Vorgehen bei der Entwicklung und dem Scoring von SJT Items eher wenig dazu geeignet ist konstrukthomogene Items hervorzubringen. Ziel der Studie liegt dementsprechend in einer Anpassung des klassischen Entwicklungsparadigmas nach Motowidlo et al. (1990) hin zu einem konstruk-

torientierten Ansatz. Dieser ist theoriegeleitet ausgerichtet und umfasst die Entwicklung konstruktrelevanter Situationen, in ihrer Traitausprägung abgestufter Antwortoptionen und der Anwendung eines konstruktorientierten Scoringverfahrens. Die theoretische Grundlage dieser Entwicklungsschritte bildet die Theorie der Trait-Aktivierung (*trait activation theory*, TAT) nach Tett und Kollegen (Tett & Burnett, 2003; Tett & Guterman, 2000; Tett, Simonet, Walser & Brown, 2013). Es werden zudem auch die Ergebnisse aus Artikel 1 berücksichtigt, um den Einfluss konstruktirrelevanter Varianz möglichst gering zu halten. Als Methode der Validierung wurde ein Multitrait-Multimethod Ansatz (Campbell & Fiske, 1959) gewählt.

2.2.2 Abstract

During the last decades there has been renewed interest in the application of Situational Judgment Tests (SJTs) both in research and practice. Even though the use of SJTs is associated with many advantages, SJTs are frequently criticized because of their largely atheoretical development and scoring process which leds to multidimensionality of SJT items. Therefore, numerous studies cast doubts on the construct validities of these low-fidelity simulations. We demonstrate that a construct-driven assessment of personality traits using SJTs is a feasible approach. We developed a SJT for the assessment of three Big Five personality traits, namely extraversion, agreeableness, and conscientiousness. Item stems and response alternatives were generated using a strictly construct-driven approach. Moreover, a theoretical scoring scheme was implemented. The newly constructed SJT and an established personality inventory (NEO-PI-R) were administered to a sample of students. Construct validity was examined by analyzing the SJT from a multitrait-multimethod perspective using different methods. The SJT exhibited convergent and discriminant validity with regard to the major dimensions of the Big Five. Moreover, the scores demonstrated adequate internal consistencies that were comparable to the reliabilities of established tools typically used in personnel selection. Contrary to the predominant negative view this study demonstrates that SJTs can exhibit construct validity. The understanding of SJTs on a conceptual level is a

crucial first step to increase their utility as a selection tool. The development of SJTs assessing specific constructs should facilitate their application in different settings and it might also lead to increased criterion validities by a better matching of predictor and criterion.

2.3 Artikel 3: SJTs und Faking

Originalpublikation: Kasten, N., Freund, P. A. & Staufenbiel, T. (eingereicht). Sweet little lies - Faking on situational judgment tests (SJT) compared to personality questionnaires.

2.3.1 Einleitende Bemerkungen

Mit dem in Artikel 2 entwickelten Verfahren ist die Erfassung der Persönlichkeitsmerkmale Gewissenhaftigkeit, Verträglichkeit und Extraversion über das SJT Paradigma möglich. Gerade bei ersterem Konstrukt besteht großes Interesse, dieses im Rahmen von Personalauswahlverfahren zu erfassen, da sich dieses als bester Prädiktor von Leistungskriterien über viele verschiede Jobs zeigt (Barrick & Mount, 1991; Ones, Dilchert, Viswesvaran & Judge, 2007). Allerdings weisen Untersuchungen, die sich mit der Verfälschbarkeit von Persönlichkeitsfragebögen beschäftigen, gerade für dieses Konstrukt ein hohes Ausmaß an Faking nach (Birkeland, Manson, Kisamore, Brannick & Smith, 2006). Aufgrund der in der Einleitung beschriebenen methodischen Schwierigkeiten von vergleichenden Studien, ist es schwer zu beurteilen, ob sich diese Verfälschungstendenzen bei SJTs in gleichem Ausmaß zeigen wie bei klassischen Persönlichkeitsfragebogen. Vor allem die fehlende Konstruktspezifikation bei den verwendeten SJTs und die damit verbundene Konfundierung zwischen Konstrukt und Messmethode machen einen Vergleich der Fakinganfälligkeit zwischen beiden Messverfahren nahezu unmöglich. Mit dem Vorliegen des in Artikel 2 entwickelten SJT ist dieses Problem obsolet. Im Rahmen von zwei empirischen Studien werden daher Unterschiede in der Fakinganfälligkeit des SJTs zu einem klassischen Persönlichkeitsfragebogen im Konstrukt

Gewissenhaftigkeit untersucht. Dazu werden, neben dem Vergleich von standardiserten Mittelwertsdifferenzen, auch verschiedene Ansätze zur Erklärung potentieller Unterschiede herangezogen, die sowohl die Antezedenzien von Faking (Studie 1) hervorheben, als auch das Vorliegen unterschiedlicher Fakingstile (Studie 2).

2.3.2 Abstract

Two laboratory studies examined potential differences in the susceptibility to faking between a construct-oriented situational judgment test (SJT) that measures conscientiousness to a traditional self-report measure of personality (NEO-FFI). In both studies standardized mean differences between honest and faked conscientiousness scores indicated that the NEO-FFI was more susceptible to faking than the SJT. In study 1 we applied a within subjects design (n = 132) and analyzed these differences in the light of different predictor variables that are based on faking models. As a result, faking on a SJT was only explained by cognitive ability whereas faking behavior on the NEO was also dependent on other personality traits that are associated with the ability to fake. In study 2 (n = 602), differences in faking susceptibility are explained by differences in faking styles. Results of mixed Rasch models indicate profound differences between measures in the way the response scale is used. Implications for researchers and practitioners are discussed.

3 Diskussion

Obgleich zahlreiche Autoren die Wichtigkeit der differenzierten Betrachtung der psychometrischen Eigenschaften von SJTs in den Vordergrund stellen (z.B. Ployhart & Ehrhart, 2003; Ryan & Ployhart, 2014; Sorrel et al., 2016), beziehen sich die meisten Studien immer noch verstärkt auf die Evaluation der Vorhersageleistung einzelner SJTs. Ausgangspunkt der vorliegenden empirischen Studien bilden daher diejenigen psychometrischen Eigenschaften von SJTs, die in bisherigen Untersuchungen keine oder wenig Beachtung gefunden haben. Dementsprechend erweitern die Ergebnisse den derzeitigen Erkenntnisstand zu SJTs in einer Reihe von Punkten. Im Rahmen der Dissertation wurden sowohl die Messgenauigkeit von SJTs im Allgemeinen als auch die Konstruktvalidität und Verfälschbarkeit eines eigens entwickelten konstruktorientierten SJTs zur Messung von Persönlichkeitskonstrukten untersucht. Viele Details wurden bereits im Rahmen der einzelnen Studien ausführlich diskutiert. Dementsprechend sollen im Rahmen der vorliegenden Diskussion nur die wichtigsten Ergebnisse der einzelnen Artikel dargestellt werden. Darüber hinaus sollen diese aber auch in Bezug zu bisherigen Forschungserkenntnissen gesetzt werden. Des Weiteren werden kritische Aspekte und Limitationen der Studien diskutiert. Die Synopsis schließt mit einem Ausblick auf Ansatzpunkte für weitere Studien.

3.1 Reliable Messungen mit SJTs

Durch die Vielzahl an Studien, die in Artikel 1 berücksichtigt wurden ist erstmals eine differenzierte Betrachtung der Reliabilität von SJTs möglich. Auf deskriptiver Ebene können dadurch Aussagen zur gängigen Publikationspraxis bezüglich Dokumentation und Auswahl von Reliabilitätskoeffizienten vorgenommen werden. Zudem wurde mit der Methodik der Reliabilitätsgeneralisierung ein sehr umfassender metaanalytischer Analyseansatz gewählt, um sowohl Aussagen zu einem mittleren Reliabilitätsniveau als auch zum Einfluss bedeutsamer Studien- und Testcharakteristika ableiten zu können. Gerade letztgenannter Punkt wurde in der bestehenden Literatur stark vernachlässigt, scheint allerdings in Hinblick auf die bestehende Heterogenität von SJT-Charakteristika sehr bedeutsam.

Es zeigte sich, dass nur in etwa 60 % der empirischen Studien überhaupt Angaben zur Messgenauigkeit des verwendeten SJTs berichtet wurden. Obwohl dieser Wert noch deutlich unter der obligatorischen Angabe von Reliabilitätskoeffizienten liegt, wie es z.B. von der APA gefordert wird (Wilkinson & American Psychological Association Task Force on Statistical Inference, 1999), ist dennoch eine positive Entwicklung im Vergleich zu der Quote von McDaniel et al. (2001) zu verzeichnen. Hier konnten nicht einmal aus der Hälfte der Studien Angaben zur Reliabilität von SJTs gezogen werden. Darüber hinaus ist diese Quote durchaus vergleichbar zu anderen Verfahren, die im Rahmen von Reliabilitätsgeneralisierungen analysiert wurden (vgl. Vacha-Haase & Thompson, 2011).

Obwohl sich unterschiedliche Quellen von Messfehlern über die verschiedenen Reliabilitätsschätzungen analysieren lassen (Nimon, Zientek, & Henson, 2012), wurden im Rahmen der empirischen SJT Studien fast ausschließlich interne Konsistenzschätzungen in Form von Cronbach's α angegeben. Nicht mal 10 % der empirischen Studien, die im Rahmen der Metaanalyse untersucht wurden, berichteten andere Koeffizienten als Cronbach's α. Diese Fokusierung auf interne Konsistenzschätzungen zeigt sich in der Forschung auch bei vielen anderen Verfahren (Vacha-Haase & Thompson, 2011) und ist vor allem damit zu begründen, dass die Berechnung dieser Reliabilitätskoeffizienten wenig Anforderungen an das Versuchs-

design richtet und sich schon bei einmaliger Testung bestimmen lässt. Allerdings ist dieses Vorgehen gerade bei SJTs kritisch zu sehen, da hier die Angemessenheit von Cronbach's α häufig in Zweifel gezogen wird (Catano et al., 2012). Das hängt vor allem damit zusammen, dass SJTs in der Regel keine homogenen Messungen unidimensionaler Konstrukte darstellen. Studien zu diesem Thema stellen heraus, dass bei bestehender Multidimensionalität die Berechnung von Cronbach's α zu einer verzerrten Reliabilitätsschätzung im Sinne einer Unterschätzung führt (Green & Yang, 2009; Kamata, Turhan & Darandari, 2003; Osburn, 2000; Zimmerman, Zumbo & Lalonde, 1993). In der SJT-Literatur zeigt sich demnach eine deutliche Divergenz zwischen der Bewertung der Angemessenheit von Cronbach's α zur Quantifizierung der Reliabilität und dessen Anwendungshäufigkeit: Obwohl viele Autoren erkennen, dass von Cronbach's α für das vorliegende SJT zu verzerrten Schätzungen führen wird, werden kaum andere Koeffizienten berichtet. Dadurch lässt sich auch erklären, dass viele α-Koeffizienten, die im Rahmen von SJT Studien berichtet werden unter den gängigen Mindestanforderungen zur Anwendung eines Messverfahrens im Forschungs- und Anwendungskontexten liegen (McDaniel & Whetzel, 2007). Die vorliegenden Ergebnisse relativieren diese überwiegend negative Sichtweise: Wie in Abbildung 2 ersichtlich wird, konnte für eine nicht unbedeutende Anzahl an SJTs (etwa ein Drittel) eine ausreichende Messgenauigkeit festgestellt werden. Darüber hinaus lag der mittlere α-Wert deutlich über den Angaben, die von Catano et al. (2012) getroffen wurde. Erstaunlich ist vor allem auch die ausgeprägte Spannweite der Reliabilitätskoeffizienten, die fast das gesamte Spektrum möglicher Schätzungen umfasste (*Range* = .00 - .93). Dementsprechend liegt der Schluss nahe, dass die mittlere Abschätzung der Messgenauigkeit kein guter Repräsentant für die Verteilung an Reliabilitätskoeffizienten ist. Das konnte auch empirisch untermauert werden. In diesem Sinne zeigten sich signifikante Anteile unerklärter Varianz auf beiden Analyseebenen. Zudem führte die Erweiterung des Modells um die Moderatorvariablen zu einem bedeutsam verbesserten Modellfit ($\Delta\chi^2$=179.847, $p < .001$).

Die Berücksichtigung der umfassenden Menge an Test- und Studiencharakteristika, die als Moderatoren in die Analyse einbezogen wurden, ermöglicht erstmals die differenzierte

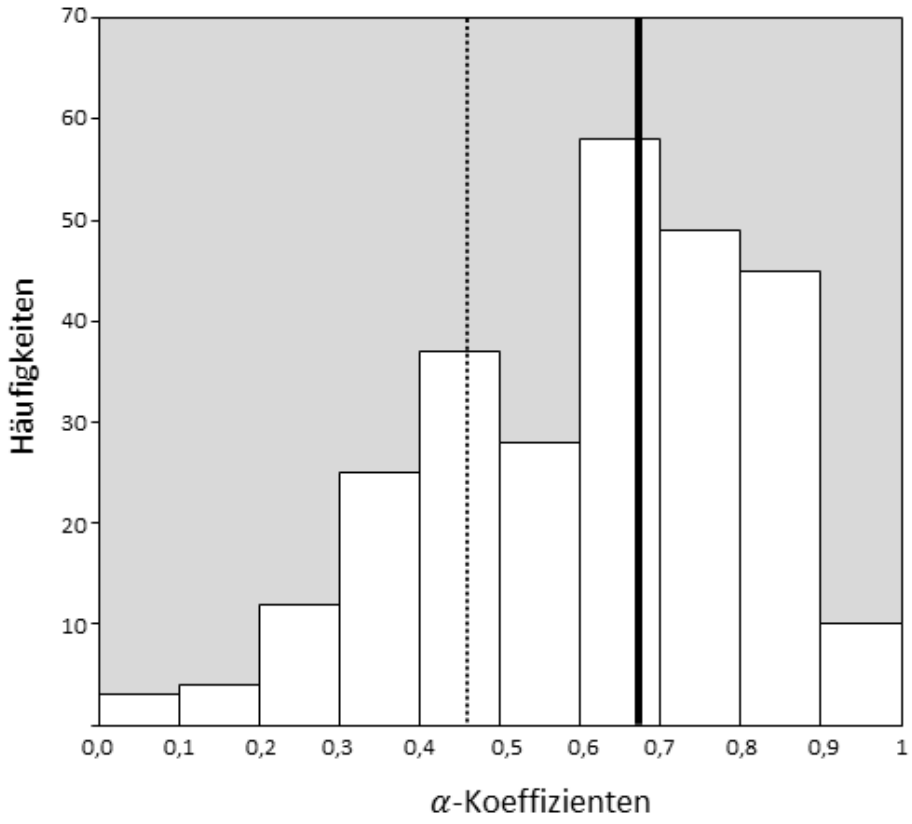

Abbildung 2. Häufigkeitsverteilung der analysierten α-Koeffizienten (k=271). Zusätzlich eingezeichnet die mittlere, gewichtete Reliabilitätsschätzung von Catano et al. (2012; gestrichelte Linie) und der vorliegenden Analyse (durchgezogene Linie).

Darstellung des Einflusses auf die Messgenauigkeit von SJTs. Signifikante Koeffizienten ergaben sich für die Testlänge, die Itemkomplexität, das Scoringverfahren, das Antwortformat, das Präsentationsmedium, den Anwendungskontext und das Publikationsjahr. Die Ergebnisse gehen größtenteils in eine erwartungskonforme Richtung, allerdings konnte kein positiver Einfluss des konstruktorientierten Vorgehens bei der Entwicklung von SJT Items auf die Messgenauigkeit dargestellt werden. Das bedeutet, dass es zu keiner Verbesserung in der internen Konsistenzschätzung kommt, wenn Items auf spezielle homogene Konstrukte hin entwickelt werden. Gegenteilig zeigt sich sogar eine tendenzielle Verschlechterung im Vergleich zum klassischen expertenbasierten Entwicklungsansatz nach Motowidlo et al. (1990). Zwar muss man festhalten, dass eine hohe interne Konsistenz nicht gleichzusetzen ist mit unidimensionalen Messungen (vgl. Green, Lissitz & Mulaik, 1977), d.h. für einzelne SJTs ist es durchaus möglich trotz bestehender Multidimensionalität hohe α-Koeffizienten zu beobach-

ten. Generell ist es aber so, dass α eine Quantifizierung der Item-Interkorrelationen darstellt und sich dementsprechend als Funktion der Multidimensionalität verringert (Cortina, 1993; Cronbach, 1951). Wie lässt sich also der fehlende Effekt für den konstruktorientierten Ansatz erklären? Bisher gibt es in den einzelnen Studien statt eines einheitlichen Paradigmas viele verschiedene Vorgehensweisen, die zudem in unterschiedlicher Weise dazu geeignet scheinen, konstrukthomogene Items hervorzubringen. Darüber hinaus scheint für die Entwicklung eines konstrukthomogenen Verfahrens auch das Scoring von entscheidender Bedeutung zu sein. Richtet man Iteminhalte und Antwortoptionen auf zuvor definierte Konstrukte aus, so wird sich nur dann ein konstrukthomogenes SJT zeigen, wenn auch die Gewichtung der Antwortoptionen anhand dieser Konstrukte vorgenommen wird. Obwohl nur bei sehr wenigen SJTs ein solcher konstruktorientierter Scoringschlüssel angewendet wurde (bei nur etwa 8 %), konnte für diesen dennoch ein positiver Einfluss auf die Messgenauigkeit im Vergleich zum expertenbasierten Ansatz nachgewiesen werden.

Die dargestellten Erkenntnisse tragen nicht nur viele theoretische Implikationen mit sich, die das Verständnis von und den Vergleich zwischen SJT Reliabilitäten fördern, sie können auch in besonderem Maße bei der Entwicklung konstrukthomogener SJTs genutzt werden. Dementsprechend bildeten sie auch die Basis für die Entwicklung des konstruktorientierten SJTs in Artikel 2. In diesem Sinne wurde bei der Entwicklung des SJTs auf eine ausreichende Itemanzahl geachtet, auf möglichst wenig komplexe Itemstämme, die schriftlich präsentiert wurden, um den Einfluss konstruktirrelevanter Varianz möglichst gering zu halten, und die Anwendung eines Rating-basierten Antwortformats. Darüber hinaus wurde das konstruktorientierte Vorgehen bei der Entwicklung und dem Scoring von SJT Items spezifiziert.

3.2 Messung spezifischer Konstrukte

Mit dem im Rahmen der Dissertation entwickelten Verfahren liegt ein konstruktvalides SJT zur Messung der Persönlichkeitskonstrukte Extraversion, Gewissenhaftigkeit und Ver-

träglichkeit vor. Nach Ausschluss nicht geeigneter Items, konnte sowohl konvergente als auch diskriminante Validität über verschiedene Multitrait-Multimethod Methodiken nachgewiesen werden. Auch die internen Konsistenzschätzungen sind für das entwickelte Verfahren als gut zu bezeichnen (α=.84 für Gewissenhaftigkeit, α=.83 für Extraversion und α=.86 für Verträglichkeit). Obwohl diese Reliabilitätsschätzungen unter denen liegen, die sich mit Hilfe des NEO-Persönlichkeitsinventar (NEO-PI-R, in der deutschen Fassung von Ostendorf & Angleitner, 2004) in Artikel 2 beobachten lassen (hier liegen die Reliabilitäten zwischen α=.93 für Verträglichkeit und α=.95 für Gewissenhaftigkeit), muss man bedenken, dass beim NEO-PI-R die Persönlichkeitskonstrukte jeweils mit mehr als dreimal so vielen Items erfasst werden. Korrigiert man für diese Unterschiede in der Itemanzahl mit Hilfe der Spearman-Brown-Formel, so ergeben sich auch für das SJT hypothetische Reliabilitäten, die über .90 liegen.

Während für das klassische Verfahren nach Motowidlo und Kollegen (1990) klare Handlungsempfehlungen bestehen, wie man bei der Entwicklung von SJT Items vorgehen soll, fehlten derartige Empfehlungen bisher gänzlich, wenn es um die Entwicklung eines konstruktorientierten SJTs geht. Eine Stärke des vorliegenden Artikels besteht dementsprechend in der Verdichtung der empirischen und theoretischen Erkenntnisse hin zu derartigen Handlungsanweisungen. Abbildung 3 verdeutlicht die Unterschiede zwischen dem klassischen und dem konstruktorientierten Vorgehen. Die verschiedenen Entwicklungsschritte sollen im folgenden kurz dargestellt und miteinander verglichen werden. Darüber hinaus soll auch diskutiert werden, inwieweit das konstruktorientierte Vorgehen bei der Entwicklung des vorliegenden SJTs zielführend war, bzw. welche Schritte noch kritisch zu sehen sind.

Konstruktspezifikation. Wie ersichtlich wird, ist das konstruktorientierte Verfahren um den Punkt der *a priori* Konstruktspezifikation erweitert worden, d.h. der vorhergehenden Festlegung auf Konstrukte, die mit dem SJT gemessen werden sollen. Während nach dem klassischen Verfahren häufig gar keine Konstruktspezifikation stattfindet oder wenn überhaupt, dann nur eine nachträgliche Zuordnung der Handlungsalternativen zu meist groben Konstruktkategorien, so ist diese der obligatorische erste Schritt bei der Entwicklung

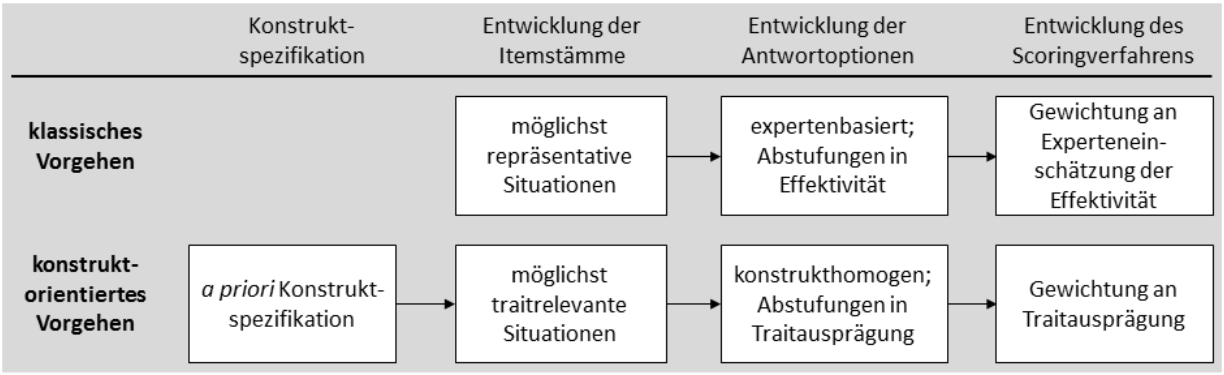

Abbildung 3. Vergleich des klassischen Entwicklungs- und Scoringparadigmas nach Motowidlo et al. (1990) und des konstruktorientierten Vorgehens

eines konstruktorientierten SJTs. Die Auswahl der Konstrukte sollte hierbei in entscheidender Weise die Brauchbarkeit eines SJTs zu Auswahlzwecken beeinflussen. Es müssen demnach solche Konstrukte angestrebt werden, denen für die vakante Stelle besondere Leistungsrelevanz zugesprochen werden (Fallgatter, 2013). Im vorliegenden Fall wurden die Persönlichkeitskonstrukte Gewissenhaftigkeit, Verträglichkeit und Extraversion ausgewählt, da sich diese in verschiedenen Kontexten als valide Prädiktoren von Leistungskriterien erwiesen haben (Barrick & Mount, 1991; Hogan & Holland, 2003; Hurtz & Donovan, 2000; Tett, Jackson & Rothstein, 1991). Während die Konstruktspezifikation eine Besonderheit des konstruktorientierten Vorgehens darstellt, orientierten sich die weiteren Entwicklungsschritte an dem dreistufigen Vorgehen des klassischen Verfahrens nach Motowidlo et al. (1990).

Entwicklung der Itemstämme. Im klassischen Ansatz besteht das Ziel bei der Entwicklung der Itemstämme vorrangig darin, Situationsbeschreibungen zu erstellen, die ein möglichst repräsentatives Bild der realen beruflichen Aufgaben und Anforderungen vermitteln. Dementsprechend werden im Rahmen dieses Vorgehens Experten (SMEs) eingesetzt. Diese bestehen zumeist aus einer Gruppe langjähriger Mitarbeiter oder Vorgesetzten, also Personen die mit den alltäglichen Arbeitsaufgaben vertraut sind. Im Gegensatz zu diesem Vorgehen ist die Entwicklung der Itemstämme beim konstruktorientierten Ansatz theoriegeleitet ausgelegt. Die theoretische Basis bildet das Konzept der Trait-Aktivierung (*trait activation theory*, TAT) nach Tett und Kollegen (Tett & Burnett, 2003; Tett & Guterman, 2000; Tett et al., 2013). Obwohl diese Theorie schon zur Spezifizierung der Konstruktvalidität von

Assessment Centern herangezogen wurde (Lievens, Chasteen, Day & Christiansen, 2006), hat die TAT bisher noch keine Beachtung bei der Entwicklung von *low-fidelity* Simulationen bekommen. Im Rahmen dieser Theorie wird postuliert, dass Persönlichkeitsmerkmale zwar individuelle Dispositionen darstellen, sich in bestimmter Weise zu verhalten, dass das Verhalten aber nicht notwendigerweise in einer Situation gezeigt werden muss. Vielmehr sind es vor allem die Merkmale der Situation, die entscheiden, ob ein bestimmtes Persönlichkeitsmerkmal gezeigt wird. Nur wenn eine Situation traitrelevante Reize aufweist, kommt es zu einer sogenannten Traitaktivierung und dementsprechend kann man nur dann von dem gezeigten Verhalten auf die zugrundeliegende Traitausprägung schließen. So kann man beispielsweise aus dem Verhalten von Personen in einer Situation, in der es um die Organisation multipler Aufgaben geht, durchaus Rückschlüsse auf deren Gewissenhaftigkeit ziehen, man erhält allerdings keine Informationen darüber, wie verträglich die Personen sind. Tett und Burnett (2003) halten dazu fest: „A situation is relevant to a trait if is thematically connected by the provision of cues, responses to which (or lack of responses to which) indicate a person's standing on the trait" (S. 502). Dementsprechend wurde bei der Entwicklung der Situationsbeschreibungen des vorliegenden SJTs darauf geachtet, dass die beschriebene Situation für das intendierte Konstrukt relevant ist. Zudem wurde darauf geachtet, dass die Situationen knapp und wenig komplex beschrieben sind. Die Intention lag hierbei vor allem darin, den Einfluss konstruktirrelevanter Varianz (z.B. über das Leseverständnis) möglichst gering zu halten.

Entwicklung der Antwortoptionen. Im klassischen Entwicklungsansatz nach Motowidlo et al. (1990) werden, ähnlich der Entwicklung der Situationsbeschreibung, SMEs eingesetzt, um die verschiedenen Antwortoptionen zu erstellen. Dadurch sollen möglichst realistische Handlungsalternativen generiert werden, die sich hinsichtlich ihrer Effektivität unterscheiden, d.h. sowohl Handlungen beinhalten, die in der vorgegebenen Situation erfolgsversprechend sind, als auch solche, die eher weniger angebracht sind. Da hier allerdings üblicherweise viele Antwortoptionen von einer Vielzahl von SMEs generiert werden, werden den Situationsbeschreibungen zumeist multidimensionale Antwortoptionen zugeordnet (Muck, 2013). Dieses

Ausmaß an Konstruktheterogenität kann sogar noch weiter gehen und auf Ebene der einzelnen Antwortoptionen angesiedelt sein, da die durch SMEs generierten Handlungsalternativen „oftmals eine Funktion multipler Eigenschaften ist“ (Muck, 2013; 187). Für die Entwicklung eines konstruktorientierten SJTs musste dieses Vorgehen entsprechend verändert werden. Ausgehend von der theoretischen Basis der TAT, lies sich auch für die Entwicklung der Antwortoptionen ein konstruktorientiertes Vorgehen ableiten. Da die Situationsbeschreibungen so entwickelt wurden, dass sie Relevanz für jeweils ein latentes Konstrukt aufwiesen, wurden die Antwortoptionen dementsprechend konstrukthomogen formuliert, d.h. nur auf ein Merkmal ausgerichtet. Abstufungen zwischen den Antwortoptionen erfolgten dementsprechend primär in Hinblick auf die Merkmalsausprägung. Für den vorliegenden Fall wurden zu jedem Itemstamm drei Antwortoptionen generiert. Eine beinhaltete Verhalten, dass für eine positive Merkmalsausprägung (*positive trait expression*, PTE) steht, also dementsprechend gewissenhafte, verträgliche oder extravertierte Verhaltensweisen beinhaltet. Eine weitere Option spiegelt jeweils eine negative Merkmalsausprägung (*negative trait expression*, NTE) wider und die dritte Antwortoption ist bezogen auf ihre Merkmalsausprägung zwischen den anderen beiden angesiedelt (*moderate trait expression*, MTE). Für das in Artikel 2 entwickelte SJT hat diese Umsetzung überwiegend gut funktioniert. Ersichtlich wird dies in Abbildung 4. Diese beinhaltet die Korrelationen der Antwortoptionen der 13 Gewissenhaftikeits-Items des SJTs[3] mit der entsprechenden Skala des NEO-PI-R. Hier wird deutlich, dass die Ratings der PTE-Antwortoptionen, die gewissenhafte Handlungen beschreiben, auch tatsächlich mittlere bis hohe positive Korrelationen mit der Gewissenhaftigkeitsskala des NEO-PI-R aufweisen. Zudem weisen die Ratings der NTE-Antwortalternativen, in fast gleichem Ausmaß, negative Korrelationen mit dem NEO-PI-R auf. Weniger einheitlich fallen die Ergebnisse jedoch für die MTE-Optionen aus. So liegen die Korrelationen, die sich für diese Antwortoptionen beobachten lassen, nicht für alle Items zwischen denen, die sich für die PTE- und die NTE-

[3]Der Übersichtlichkeit halber werden hier nur die Ergebnisse der Gewissenhaftigkeitsskala berichtet. Da die Ergebnisse sowohl für Verträglicheit als auch für Extraversion ähnlich sind, wird an dieser Stelle nicht mehr gesondert darauf eingegangen. Die entsprechenden Abbildungen finden sich aber im Anhang

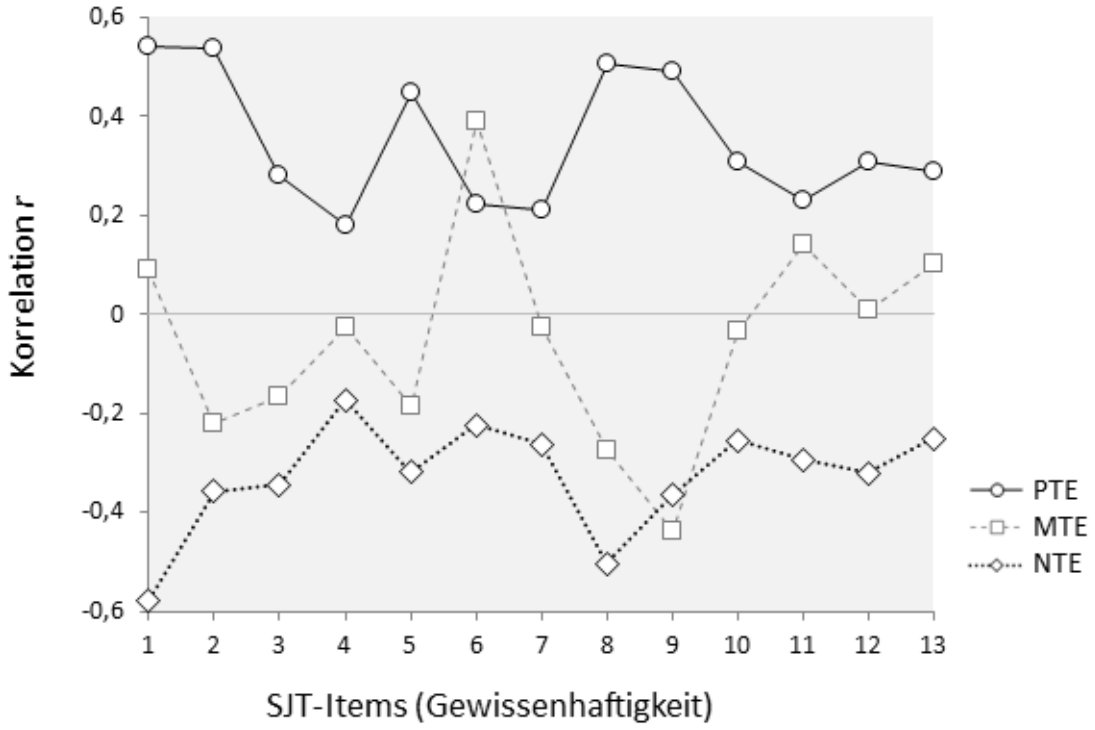

Abbildung 4. Korrelationen der Antwortoptionen der 13 Items der SJT-Gewissenhaftigkeitsskala mit der entsprechenden Skala des NEO-PI-R. PTE = *positive trait expression*, MTE = *moderate trait expression*, NTE = *negative trait expression*. Berechnungen anhand des Datensatzes aus Artikel 2 ($N = 162$)

Optionen ergeben. Zudem ist das Zusammenhangsmuster nicht einheitlich, d.h. es finden sich sowohl positive als auch negative Korrelationen mit der Gewissenhaftigkeitsskala des NEO-PI-R.

Entwicklung des Scoringverfahrens. Im Rahmen des klassischen Verfahrens werden, wie schon bei den vorangegangenen Entwicklungsschritten, auch bei der Entwicklung eines Scoringschlüssels Experten eingesetzt. Diese werden gebeten die Effektivität der einzelnen Handlungsalternativen einzuschätzen. Obwohl das spezifische Vorgehen hier durchaus unterschiedlich ausfällt, wird in den zumeist noch ein Forced-Choice Format eingesetzt, d.h. die Experten sollen sowohl die beste, als auch die schlechteste Handlungsalternative in der gegebenen Situation benennen. Da sich im konstruktorientierten Fall die Unterschiedlichkeit der Antwortoptionen zunächst einmal nicht auf die Effektivität der dargestellten Handlungen bezieht, sondern auf den Grad der Merkmalsausprägung, wurde dies auch bei der Entwick-

lung des Scoring-Schlüssels berücksichtigt. Konkret wurde für jedes Item ein Mittelwert aus den PTE-Antworten und den invertierten NTE-Antworten gebildet. Die MTE-Antworten wurden beim Scoring nicht berücksichtigt, da sich hier kein einheitliches Korrelationsmuster ergab (siehe Abbildung 4).

Obwohl über die dargestellte Anpassung des Entwicklung- und Scoringansatzes im vorliegenden Fall ein konstruktvalides SJT entwickelt werden konnte, müssen noch weitere Studien folgen, um die Zweckmäßigkeit dieses Vorgehens zu untermauern. Das hängt unter anderen mit den Limitationen der Studie zusammen. So ist unklar, ob das dargestellte konstruktorientierte Vorgehen auch auf andere Konstrukte übertragbar ist. Die Antwortoptionen eines SJTs stellen in der Regel - und da bildet das hier dargestellte SJT keine Ausnahme - Handlungsalternativen dar und sind dementsprechend verhaltens-verankert. Im vorliegenden Fall wurden daher explizit die Persönlichkeitsdimensionen ausgewählt, deren Exposition sich auch direkt im Verhalten von Personen zeigt. Dementsprechend wurden hier nur Items für die Persönlichkeitskonstrukte Extraversion, Gewissenhaftigkeit und Verträglichkeit entwickelt, nicht aber für Neurotizismus und Offenheit. Zwar äußern sich auch diese indirekt im Verhalten, allerdings sind sie per Definition viel stärker an internale Emotionen und Einstellungen geknüpft (McCrae & Costa, 2008). Die Entscheidung, sich auf die verhaltens-verankerten Dimensionen zu beschränken, wurde auch in Hinblick auf bisherige empirische Studien getroffen, bei denen vor allem die internal-verankerten Dimensionen Probleme bei der Konstruktvalidierung oder niedrige α-Koeffizienten aufwiesen, nicht aber verhaltens-verankerte Dimensionen (vgl. z.B. Ahnefeld, 2009; Trippe, 2002). Für zukünftige SJT-Entwicklungen wäre es aber durchaus von Interesse zu untersuchen, ob der hier dargestellte Entwicklungs- und Scoringansatz auch auf solche internal-verankerten Eigenschaften übertragbar ist, oder ob sich bedeutsame Unterschiede zeigen.

Weitere Limitationen der Studie liegen in den fehlenden Hinweisen zur situations- und personenbezogenen Generalisierbarkeit des entwickelten SJTs begründet. Das hängt vor allem damit zusammen, dass sich die Stichprobe aus Studierenden zusammensetzte und die Untersuchung in einem Laborsetting stattfand. Auf Grundlage bestehender empirischer Stu-

dien lässt sich allerdings vermuten, dass bei Anwendung in einem realen Auswahlsetting durchaus divergente Ergebnisse für das SJT zu erwarten sind, z.B. in Form veränderter faktorieller Strukturen (Klehe et al., 2012; Schmit & Ryan, 1993) oder niedrigerer Reliabilitätsschätzungen (Kasten & Freund, 2016). Wünschenswert wäre dementsprechend die Überprüfung der Ergebnisse anhand von unabhängigen und divergenten Stichproben, z.B. im Sinne einer Kreuzvalidierung.

3.3 Verfälschbarkeit von SJTs

Das in Artikel 2 entwickelte konstruktvalide SJT ermöglicht erstmalig einen direkten Vergleich zwischen der Fakinganfälligkeit von SJTs und einem klassischen Persönlichkeitsfragebogen, dem NEO-Fünf-Faktoren-Inventar (NEO-FFI; deutsche Version von Borkenau & Ostendorf, 2008). Es zeigt sich in zwei Studien, dass das SJT in deutlich geringerem Ausmaß Antwortverfälschungen aufweist, als das NEO-FFI. Allerdings muss man festhalten, dass die vorliegenden Ergebnisse im Vergleich zu anderen Studien, die sich mit der Verfälschbarkeit von SJTs befassen, deutlich weniger positiv ausfallen, da sich auch für das SJT in beiden Studien deutliche Antwortverfälschungen zeigten. In Studie 1 von Artikel 3 liegen die standardisiersten Mittelwertsunterschiede zwischen den beiden Instruktionsformen (Normal vs. Faking) für den SJT bei $d = 1.06$ und $d = 0.84$, in Studie 2 bei $d = 0.62$ und $d = 0.66$. Doch auch wenn diese Werte über den Effektstärken liegen, die sich typischerweise mit SJTs beobachten lassen (vgl. Hooper et al., 2006), so sind sie dennoch vor dem Hintergrund der bestehenden Fakingliteratur erklärbar. Erstens wurde im vorliegenden Fall eine verhaltensbasierte, also eine „*would do*“-Instruktion für das SJT gewählt. Vor dem Hintergrund, dass das Ziel darin bestand Persönlichkeitskonstrukte zu erfassen, ist dieser Schritt naheliegend. Allerdings sind bei diesem Instruktionstyp Antwortverfälschungen wahrscheinlich, da typisches Verhalten erfasst werden soll. Demgegenüber sollte Faking bei wissensbasierten Instruktionen keine Rolle spielen, da hier ja bereits maximales Verhalten erfasst wird (Chan & Schmitt, 2005). Ein Vergleich zu anderen SJTs kann also dementsprechend nur

unter Berücksichtigung des Instruktionstyps gezogen werden.

Zweitens bestimmt sich das Ausmaß von Antwortverfälschungen nicht nur über die verwendete Instruktion, sondern auch in nicht unerheblichen Maße über andere Charakteristika der Testitems (Snell, Sydell & Lueke, 1999). Demnach sollten Items vor allem dann einfacher zu verfälschen sein, je transparenter diese für den Bewerber sind. Transparenz bestimmt sich nach dieser Sichtweise vor allem durch die Leichtigkeit, mit der Probanden Hypothesen über die zugrundeliegenden Traits bilden und auf dieser Grundlage adäquate von inadäquaten Verhaltensweisen unterscheiden können (Dilchert & Ones, 2012; Snell et al., 1999). Für das vorliegende SJT sind es wahrscheinlich vor allem der Scoring-Key, das Antwortformat und die geringe Itemkomplexität, die im Vergleich zu anderen SJTs zu einer ausgeprägteren Transparenz und dementsprechend zu einer erhöhten Fälschungsanfälligkeit führen.

Die Ergebnisse aus Artikel 3 erlauben, neben dem Vergleich im generellen Ausmaß an Antwortverfälschungen, zudem eine Einordnung in die bestehende Faking-Literatur, die sich bisher zum Großteil nur auf klassische Selbstbeurteilungsfragebogen der Persönlichkeit stützt. In Studie 1 wurde der Fokus auf die Antezedenzien von Faking gelegt. Es wurden dabei, unter Rückgriff auf etablierte Fakingmodelle (McFarland & Ryan, 2000; 2006; Mueller-Hanson, Heggestad & Thornton, 2006; Snell et al., 1999) Dispositionen erhoben, die sowohl mit der Fähigkeit zu faken als auch mit der Motivation, dies zu tun, zusammenhängen. Hier zeigten sich Überscheidungen zwischen den beiden Messverfahren, in dem Sinne, dass bei beiden die kognitiven Fähigkeiten einer Person einen signifikanten Einfluss auf die Verfälschung der Gewissenhaftigkeitsausprägung hatte. Für das SJT war der Einfluss der Intelligenz aber ausgeprägter und zudem der einzig signifikante Prädiktor. Dieser Unterschied steht im Einklang mit der bisherigen Vermutung, dass SJTs vor allem deswegen weniger fälschungsanfällig sind, da sie aufgrund erhöhter Komplexität durch die genestete Itemsstruktur, höhere Anforderungen an die kognitiven Fähigkeiten der Bewerber richten (Hooper et al., 2006).

In Studie 2 aus Artikel 3 wurden zudem die Unterschiede im Fakingverhalten zwischen den Messinstrumenten auf Itemebene untersucht. Dazu wurden mit Hilfe von Mixed-Rasch

Analysen qualitative Unterschiede im Fakingstil der Personen untersucht. Auch hier zeigten sich Ähnlichkeiten zwischen den Verfahren. Für beide Verfahren ließen sich drei verschiedene Fakingstile unterscheiden, die sich so auch in anderen Studien separieren lassen (Eid & Zickar, 2007; Zickar, Gibby & Robie, 2004, Ziegler & Kemper, 2013): Eine Gruppe von Probanden, die einen extremen Fakingstil aufweist, eine Gruppe, die nur in geringem Maße von ihren tatsächlichen Gewissenhaftigkeitswerten abweicht, und solche Probanden, die trotz der Fakinginstruktion die Items ehrlich beantworten. Unterschiede zwischen den Messverfahren manifestierten sich vor allem über die Verteilung der Probanden auf die verschiedenen Gruppen und über quantitative Unterschiede im Fakingausmaß innerhalb der Gruppen. So ist das Ausmaß an Antwortverfälschungen in allen Gruppen im SJT jeweils deutlich geringer ausgeprägt als bei der entsprechenden Gruppe im NEO-FFI.

Eine Limitation der Studien liegt auch wie bei Artikel 2 in der zweifelhaften Generalisierbarbeit der Ergebnisse. Zwar wurde hier auf die Realisation heterogenerer Stichpoben geachtet, indem zumindest in Studie 2 neben Studierenden auch berufstätige Probanden miteinbezogen wurden, allerdings wurde in beiden Studien von Artikel 3 ein experimentelles Labordesign angewendet. Für natürliche Settings im Rahmen von tatsächlichen Auswahlverfahren sind divergente Effekte wahrscheinlich und teilweise empirisch nachweisbar (vgl. MacCann, Ziegler & Roberts, 2012). Allerdings ist festzuhalten, dass sich in Laborstudien, bei denen Personen explizit dazu aufgefordert werden sich möglichst positiv darzustellen, Faking häufig in einem sehr viel stärkeren Ausmaß gezeigt wird, als im Rahmen von natürlichen Bewerbersettings. Untersucht wird im Laborsetting dementsprechend, ob das SJT überhaupt anfällig gegenüber Verfälschungen ist, nicht aber, ob tatsächliche Bewerber es in einem Selektionskontext auch wirklich verfälschen. In letzterem Fall ist von einer sehr viel stärkeren Heterogenität bezüglich der individuellen Motivation zu Fälschen auszugehen. Dementsprechend stellen unsere Ergebnisse Worst-Case-Szenarien dar.

3.4 Ausblick

Über die bisher diskutierten Erkenntnisse hinausgehend sollen im folgenden die Aspekte näher betrachtet werden, auf die die vorliegenden Studien keine Antwort liefern und die dementsprechend in weiteren Studien untersucht werden sollten. Obwohl auch in den vorangegangenen Abschnitten schon auf verschiedene Ausgangspunkte für Folgeuntersuchungen hingewiesen wurde, sollen hier vor allem zwei Punkte diskutiert werden, die meines Erachtens nach von besonderer Bedeutung sind. Ersterer betrifft die Frage, nach der Anwendbarkeit konstruktorientierter SJT in realen Auswahlsetting. Hier sollen vor allem die Fragen diskutiert werden, die noch untersucht werden müssen, um die Angemessenheit der Anwendung konstruktorientierter SJTs zur Bewerberauswahl einschätzen zu können. Zweitens soll betrachtet werden, welchen Mehrwert die dargestellten Erkenntnisse bezüglich einer derzeit geführten Diskussion zur Rekonzeptualisierung von SJTs haben. Hier stehen vor allem die Fragen im Vordergrund, wie man das hier entwickelte SJT nutzen kann, um das theoretische Verständnis von SJTs zu vertiefen.

3.4.1 Konstruktorientierte SJTs als Instrumente der Personalauswahl

Die hier dargestellten Erkenntnisse erlauben noch zu wenig Einsicht, um eine Anwendung des entwickelten konstruktorientierten SJTs in solchen Kontexten zu rechtfertigen, in denen auf Grundlage der Testergebnisse wichtige Auswahlentscheidungen getroffen werden (sogenannte *high-stakes* Kontexte). Dementsprechend müssen noch weitere Studien folgen, die diejenigen psychometrischen Eigenschaften des konstruktorientierten SJTs untersuchen, die im Rahmen der Personalauswahl von besonderer Bedeutung sind. Bevor im einzelnen auf die relevanten Eigenschaften eingegangen wird, muss angemerkt werden, dass die Grundlage der Evaluation weniger die Frage sein kann, wie das konstruktorientierte SJT im Vergleich zu anderen SJTs abschneidet, sondern inwiefern es eine alternative Methode zur Erfassung der Persönlichkeit im Rahmen von Personalauswahlverfahren darstellt. Obwohl der Vergleich

zu anderen SJTs naheliegend erscheint und in der Literatur auch im Zusammenhang mit der Evaluation alternativer Entwicklungsparadigmen von SJT Items gefordert wird (Borneman, 2016), bleibt die Frage, wie aufschlussreich eine solcher Vergleich zwischen verschiedenen SJTs sein kann. Dieser Zweifel lässt sich aus der grundlegenden Differenzierung von Prädiktorkonstrukten und -methoden ableiten (Arthur & Villado, 2008; Binning & Barrett, 1989; Champbell & Fiske, 1959). Diese Unterscheidung unterliegt der Annahme, dass sich verschiedene Auswahlverfahren sowohl in den erfassten Konstrukten als auch in der spezifischen Methodik unterscheiden können. Vergleiche von psychometrischen Eigenschaften verschiedener Auswahlverfahren sind nach Arthur und Villado (2008) nur innerhalb einer konstanten Konstruktdimension oder innerhalb einer festen Methodik interpretierbar. Bei gleichzeitiger Veränderung in beiden Dimensionen sind z.B. Unterschiede in Validitätskoeffizienten nicht mehr interpretierbar, da nicht differenziert werden kann, ob diese auf Unterschiede in der Methodik oder Unterschiede in den erfassten Konstrukten zurückzuführen sind. Ein Vergleich der psychometrischen Eigenschaften des hier entwickelten SJTs zu anderen Verfahren der Persönlichkeitsmessung ist daher ohne weiteres möglich, da hier die Konstruktdimensionen konstant gehalten werden. Da bei dem Vergleich zu klassischen SJTs diese Konstanz, aufgrund der meist fehlenden Konstruktvalidität, nicht unterstellt werden kann, wäre Unterschiede in den psychometrischen Eigenschaften nur deutbar, wenn man SJTs als eine feste Methodik auffasst. Aufgrund der vielfach beschriebenen Heterogenität einer Vielzahl von Testcharakteristika scheint dieser Ansatz allerdings zumindest fragwürdig.

Bezogen auf die Kriteriumsvalidität des hier entwickelten SJTs ist mit einem Vorteil gegenüber klassischen Selbstbeurteilungsfragebögen zur Messung der Persönlichkeit zu rechnen. Ableiten lässt sich diese Hypothese aus der bestehenden Literatur zum Einfluss des Referenzrahmens (*frame-of-refence*, FOR; Schmit, Ryan, Stierwalt & Powell, 1995) auf die Vorhersageleistung von Persönlichkeitsfragebögen. Die Hauptannahme dieser Theorie besteht darin, „that prediction of people's behavior can be inproved when people are given a context, or frame-of-reference, when asked to describe themselves“ (Lievens, De Corte & Schollaert, 2008; S. 468). Die meisten Persönlichkeitsinventare geben diesen Kontext aber gerade nicht,

d.h. sie sind so entwickelt, dass sie nach generellen, situationsübergreifenden Neigungen fragen. Empirisch zeigt sich allerdings deutliche Evidenz, dass die Verwendung einer FOR, also die Erfassung situationsspezifischer statt genreller Neigungen, über die Verringerung von Fehlervarianz zu einer deutlichen Verbesserung der Kriteriumsvalidität führt (Hunthausen, Truxillo, Bauer & Hammer, 2003; Lievens, De Corte et al., 2008; Robie, Schmit, Ryan & Zickar, 2000; Schmit et al., 1995; Shaffer & Postlethwaite, 2012). Erzeugt wird dieser Referenzrahmen in empirischen Studien zumeist über die Umformulierung der Standardinstruktion von Persönlichkeitsfragebogen oder um die Erweiterung der Items um eine situationsspezifische Plakete (z.B. dem Zusatz „*bei der Arbeit*“ hinter jedem Item). Es lässt sich vermuten, dass die Kontextualisierung, die über die Situationsbeschreibung bei SJTs erzeugt wird, zumindest einen vergleichbaren Referenzrahmen herstellt. Um den Mehraufwand zu rechtfertigen, der mit der Entwicklung eines konstruktorientierten SJTs einhergeht, sollten sich aber auch darüber hinaus noch positive Effekte zeigen. Erste empirische Evidenz, dass dies der Fall sein könnte findet sich z.B. bei Holtrop, Born, de Vries und de Vries (2014). Hier zeigte ein Persönlichkeitsfragebogen, dessen Kontextualisierung über die reine Bereitstellung eines Referenzrahmens hinausging, auch darüber hinaus noch positive Effekte bezüglich der Prädiktionsleistung. Ob diese Erkenntnisse allerdings auch auf SJTs übertragbar sind, kann nur eine vergleichende Studie ermitteln, die die Kriteriumsvalidität aller drei Messverfahren (SJT vs. Persönlichkeitsfragebogen mit Standardinstruktion vs. Persönlichkeitsfragebogen mit FOM-Instruktion) evaluiert und vergleicht.

Aufgrund der in der Einleitung beschriebenen, weitreichenden Konsequenzen, die sich aus den subjektiven Bewerberwahrnehmungen des Auswahlprozesses und der -instrumente ergeben, tut ein Unternehmen gut daran, diese bei der Auswahl eines geeigneten Personalauswahlverfahrens zu berücksichtigen. Dementsprechend sollte der Nachweis angemessener Anwenderreaktionen für das vorliegende SJT Ausgangspunkt weiterer Untersuchungen sein. Auch hier ist der Vergleich zu klassischen Persönlichkeitsfragebögen besonders interessant, da sich für diese in der Literatur häufig keine positiven Reaktionen feststellen lassen (Hausknecht et al., 2004; McFarland, 2013). Auf Grundlage bestehender Modelle der Antezen-

denzien von Anwenderreaktionen (Gilliland, 1993, 1994) lässt sich vermuten, dass SJTs hier über den wahrgenommenen Berufsbezug durch die Darstellung berufsrelevanter Situationen von Vorteil sind (Kanning et al., 2006; Richman-Hirsch et al., 2000). Allerdings ist bisher noch vollkommen unklar wie spezifisch und ausführlich die Situationsbeschreibung ausfallen muss, um diesen Berufsbezug bei den Bewerbern zu triggern. Für das vorliegende SJT ist diese Frage in besonderem Maße relevant, da hier die Situationsbeschreibungen knapp und eher berufsunspezifisch formuliert sind. Dadurch sollte der Einfluss konstruktirrelevanter Varianz möglichst gering gehalten und eine generalisierte Anwendung in vielen verschiedenen Kontexten gewährleistet werden. Aus der bestehendem FOM-Literatur lässt sich auch für solche geringen Kontextualisierungen ein Vorteil ableiten, allerdings ist die empirische Evidenz nicht einheitlich (Holtrop et al., 2014; Holtz, Ployhart & Dominguez, 2005) und bedarf weiterer Untersuchungen.

3.4.2 Konzeptualisierung von SJTs

Bezogen auf die Vorstellungen darüber, was SJTs messen und warum mit ihnen Leistungsvorhersagen möglich sind, hat es in den letzten Jahren große theoretische Entwicklungen und fruchtbare Diskussionen gegeben. Auch für das im Rahmen der Dissertation entwickelte SJT bzw. das hier dargestellte konstruktorientierte Entwicklungsparadigma ergeben sich vor dem Hintergrund dieser Diskussion weitere offene Fragen, die in Folgestudien untersucht werden können. Die Diskussion, betrifft dabei die Frage, welche Bedeutung dem Itemstamm von SJTs beizumessen ist. Obwohl die Situationskomponente bzw. deren Bewertung bei SJTs sogar Bestandteil des Namens ist und die dadurch unterstellte kontextualisierte Messung als wichtiger definitorischer Aspekt angesehen wird, wurde die Bedeutsamkeit des Itemstamms vor allem in den letzten Jahren stark in Zweifel gezogen. Ausschlaggebend war vor allem der Artikel von Krumm, Lievens, Hüffmeier, Lipnevich Bendels und Hertel (2015). In mehreren Studien untersuchten die Autoren Unterschiede in der Antworteffektivität zwischen Probanden, die ein SJT unter normalen Voraussetzungen beantworteten und solchen, die

die Handlungsalternativen ohne vorangegangene Situationsbeschreibung beantworten sollten. Die Varianzanalysen zeigten keinen Effekt für diese Manipulation, d.h. es zeigte sich kein Unterschied in den Effektivitätseinschätzungen der Handlungsalternativen zwischen den zwei Testversionen. Auch Untersuchungen auf Itemebene zeigten nur für wenige Items einen signifikanten Unterschied zwischen den Bedingungen. Die Autoren folgerten aus diesen Ergebnissen, dass SJTs bestimmtes dekontextualisiertes Wissen erfassen, sogenanntes *General Domain Knowledge* (GDK), das definiert ist als „*general rules about the utility of behavioral acts across a wide range of situations in a specic domain*“ (Krumm et al., 2015; S. 400). Diese Wissenskomponente sollte demnach relativ unbeeinflusst sein durch die Spezifika der Situation.

Diese theoretische Vorstellung von SJTs hat in den letzten zwei Jahren viel Diskussion angeregt und ist nicht unkritisiert geblieben. Ohne das zum damaligen Zeitpunkt zu ahnen[4], wurde auch im Rahmen dieser Arbeit, bei der Spezifikation des konstruktorientierten Entwicklungsparadigmas, eine konträre Sichtweise angenommen. Durch die Bindung an die TAT als theoretisches Grundgerüst zur Entwicklung konstrukthomogener SJT-Items wird ja gerade die Bedeutsamkeit der Situationskomponente betont. Zudem konnte über dieses Verfahren ein konstruktvalides SJT zu Messung von Persönlichkeitskonstrukten und eben nicht GDK entwickeln lassen konnte. Allerdings können und sollen die Ergebnisse aus Artikel 2 nicht als Nachweis gegen die dekontextualisierte Sichtweise verstanden werden. Dennoch ergeben sich vor dem Hintergrund der derzeit geführten Diskussion für das vorliegende SJT noch offene Fragen.

Fraglich ist zunächst, inwieweit die dargestellte Theorie der GDK überhaupt auf das vorliegende SJT übertragbar ist, da für dieses eine verhaltensbasierte Instruktion verwendet wurde. Da solche Instruktionstypen darauf abzielen typisches Verhalten zu erfassen, erscheint es zunächst abwegig, dass sie Wissenskomponenten in Form von GDK erfassen. In

[4]Die konzeptuellen Überlegungen, die Entwicklung des SJTs und weiteren Arbeiten an Artikel 2 fanden vorrangig 2013, also zeitlich vor der hier dargestellten Diskussion statt

der bestehenden Literatur zu GDK wird allerdings zwischen den beiden Instruktionstypen, zumindest im Rahmen von *high-stakes* Kontexten, nicht differenziert. Argumentiert wird dabei so, dass Bewerber in Auswahlkontexten bei verhaltensbasierten Instruktionen ein starkes Ausmaß an Antwortverfälschungen aufweisen und dementsprechend die Handlungsoptionen nicht gemäß der eigenen Präferenz, sondern entgegen der Instruktion nach der Effektivität bewerten (Lievens et al., 2009; Motowidlo et al., 2006; Muck, 2013). Nach dieser Argumentation wäre die Unterscheidung der Instruktionstypen in Auswahlkontexten obsolet, da beide gleiches Verhalten bei den Bewerbern erzeugen. Die im Rahmen der Dissertation dargestellte interindividuelle Heterogenität im Fakingverhalten (Artikel 3) lässt jedoch Zweifel an dieser Hypothese offen. Aufschluss zu dieser Frage könnte jedoch die Manipulation der Instruktion in einem realen Auswahlkontext liefern. Demnach würden Bewerber das SJT entweder mit einer wissensbasierten oder der bisher verwendeten verhaltensbasierten Instruktion bekommen. Ein vergleichbares Antwortverhalten sollte sich sowohl in vergleichbaren Mittelwerten und Standardabweichungen, aber auch in einer analogen Verteilung von Antwortstilen zwischen den Gruppen widerspiegeln.

Ein weiteres, vielleicht schwerwiegenderes Problem zur Klärung, was SJTs messen liegt in dem von Lievens und Motowidlo (2016) vorgeschlagenen Entwicklungsprozedere für SJTs zur Erfassung von GDK. Für die Entwicklung der Handlungsoptionen halten die Autoren fest:

> In the traditional multiple-response format, our approach requires that response alternatives vary in more than what level of effectiveness they represent. They should be effective *and* represent a high level of the targeted trait or they should be ineffective *and* represent a low level of the trait. (S. 13)

Probanden wird dann GDK attestiert, wenn sie solche Antwortoptionen, die effektiver sind, auch als solche erkennen. Da die Effektivität und die Traitausprägung in diesem Vorgehen jedoch komplett kovariieren, ist es gar nicht möglich zu differenzieren, ob mit dem SJT Persönlichkeit oder GDK erfasst wird, da in beiden Fällen ein gleiches Antwortmus-

ter zu erwarten ist. Um diese zwei Faktoren auseinander zu ziehen bietet sich ein 2x2 Design an, dass eine vollständige Faktorstufenkombination der Traitausprägung (hohe vs. niedrige Traitausprägung) und der Effektivität (effektiv vs. ineffektiv) der Handlungsoption berücksichtigt. Demnach müssten neben den bisher beschriebenen Handlungsoptionen auch solche generiert werden, die eine hohe Effektivität *und* niedrige Traitausprägung aufweisen bzw. solche, die ineffektiv sind und gleichzeitig eine hohe Traitausprägung aufweisen. Vielleicht könnte auch schon das hier entwickelte SJT verwendet werden, wenn zusätzlich noch Experteneinschätzungen der Effektivität für die einzelnen Antwortoptionen erhoben werden und sich die entsprechenden Faktorstufenkombinationen abbilden. Wird mit SJTs die Wissenskomponente GDK erfasst so sollten sich die Variabilität in den Daten vor allem auf Unterschiede in der Effektivität zurückzuführen sein, bei einer Erfassung von Persönlichkeit, eher auf Unterschiede in der Traitausprägung. Wie ausgeprägt solche Effekte allerdings sind und ob sie sich überhaupt so finden lassen, ist schwer zu prognostizieren, da auch GDK nicht unabhängig von der Persönlichkeit einer Person ist (Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010; Motowidlo, Hooper & Jackson, 2006) und der Einfluss der Traitausprägung auf die Einschätzung des Probanden durch Antworttendenzen im Sinne von Faking verzerrt werden kann.

4 Literatur

Achouri, C. (2010). *Recruiting and Placement - Methoden und Instrumente der Personalauswahl und -platzierung* (2. Auflage). Wiesbaden: Gabler.

Ahnefeld, K. (2009). *The development and validation of a nonviolent communication situational judgment test (NVC-SJT) for the workplace* (Master's thesis). Verfügbar unter: http://pqdtopen.proquest.com

Arthur, W. & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93,* 435-442.

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

Bauer, T. N. & Truxillo, D. M. (2006). Applicant reactions to situational judgment tests: Research and related practical issues. In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests: Theory, measurement and application* (S. 233-249). Mahwah: Erlbaum.

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J.M., Ferrara, P. & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology, 54,* 387-419.

Behrmann, M. (2007). Situational judgment tests. In H. Schuler & K. Sonntag (Hrsg.), *Handbuch der Arbeits- und Organisationspsychologie* (S. 483-489). Göttingen: Hogrefe.

Behrman, M. (2011). Verhandeln, Persönlichkeit und ihre Messung in Situational Judgment Tests. In P. Gelléri & C. Winter (Hrsg.), *Potenziale der Personalpsychologie. Einfluss personaldiagnostischer Maßnahmen auf den Berufs- und Unternehmenserfolg* (S. 281-293). Göttingen: Hogrefe.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B. & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14,* 223-235.

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.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T. & Smith, M. A. (2006). A meta-analytical investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14,* 317-335.

Borneman, M. J. (2016). Further considerations in SJT development. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 9,* 55-59.

Borkenau, P. & Ostendorf, F. (2008). *NEO-FFI: NEO-Fünf-Faktoren Inventar nach Costa und McCrae.* Göttingen: Hogrefe.

Carless, S. A. (2003). A longitudinal study of applicant reactions to multiple selection procedures and job and organizational characteristics. *International Journal of Selection and Assessment, 4,* 345-351.

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

Campion, M. C., Ployhart, R. E. & MacKenzie, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance, 27,* 283-310.

Catano, V. M., Brochu, A. & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment, 20,* 334-346.

Chan, D. & Schmitt, N. (2005). Situational judgment tests. In A. Evers, N. Anderson & O. Voskuijl (Hrsg.), *The Blackwell handbook of personnel selection* (S. 219-242). Malden: Blackwell Publishing.

Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A. & Jones, D. A. (2005). Applicant attraction to organizations and job choice: A meta-analytical review of the correlates of recruiting outcomes. *Journal of Applied Psychology, 90,* 928-944.

Christian, M. S., Edwards, B. D. & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63,* 83-117.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Dilchert, S. & Ones, D. S. (2012). Application of preventive strategies. In M. Ziegler, C. MacCann & R. D. Roberts (Hrsg.), *New perspectives on faking in personality assessment* (S. 177-201). New York: Oxford University Press.

Eid, M. & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessment by mixed Rasch models. In M. von Davier & C. H. Carstensen (Hrsg.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (S. 255-270). New York: Springer.

Fallgatter, M. J. (2013). Personalbeurteilung. In R. Stock-Homburg (Hrsg.), *Handbuch strategisches Personalmanagement* (S. 171-186). Wiesbaden: Springer-Gabler.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 54,* 327-358.

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

Gilliland, S. W. (1994). Effects of procedural and distributive justice on reactions to a selection system. *Journal of Applied Psychology, 79,* 691-701.

Görlich, Y. & Schuler, H. (2014). Personalentscheidung, Nutzen und Fairness. In H. Schuler & U. P. Kanning (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 1137-1199). Göttingen: Hogrefe.

Green, S. B., Lissitz, R. W. & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37,* 827-837.

Green, S. B. & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrica, 74,* 121-135.

Hauenstein, N. M. A., Findlay, R. A. & McDonald, D. P. (2010). Using situational judgment tests to assess training effectiveness: Lessons learned evaluating military equal opportunity advisor trainees. *Military Psychology, 22,* 262-281.

Hausknecht, J. P., Day, D. V. & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57,* 639-683.

Hedlund, J. & Sternberg, R. (2000). Practical intelligence: Implications for human resources research. *Research in Personnel and Human Resources Management, 19,* 1-52.

Hogan, J. & Holland, B. (2003). Using theory to evaluate personality and job performance relations: A sociaanalytic perspective. *Journal of Applied Psychology, 88,* 100-112.

Holtrop, D., Born, M. P., de Vries, A. & de Vries, R. E. (2014). A matter of context: A comparison of two types of contextualized personality measures. *Personality and Individual Differences, 68,* 234-240.

Holtz, B. C., Ployhart, R. E. & Dominguez, A. (2005). Testing the rules of justice: The effects of frame-of-reference and pretest validity information on personality test responses and test perceptions. *International Journal of Selection and Assessment, 13,* 75-86.

Hooper, A. C., Cullen, M. J. & Sackett, P. R. (2006). Operational threats to the use of SJTs: Faking, Coaching, and retesting issues. In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests: Theory, measurement and application* (S. 205-232). Mahwah: Erlbaum.

Hurtz, G. M. & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85,* 869-879.

Kamata, A., Turhan, A. & Darandari, E. (2003, April). *Estimating reliability for multidimensional composite scale scores.* Paper presented at the Annual Meeting of American Educational Research Association, Chicago, IL.

Kanning, U. P. (2015). *Personalauswahl zwischen Anspruch und Wirklichkeit: Eine wirtschaftspsychologische Analyse.* Berlin: Springer.

Kanning, U. P., Grewe, K., Hollenberg, S. & Hadouch, M. (2006). From the subject's point of view: Reactions to different types of situational judgment items. *European Journal of Psychological Assessment, 22,* 168-176.

Hunthausen, J. M., Truxillo, D. M., Bauer, T. N. & Hammer, L. B. (2003). A field study of frame-of-reference effects on personality test validity. *Journal of Applied Psychology, 88,* 545-551.

Kanning, U. P. & Kuhne, S. (2006). Social desirability in a multimodal personnel selection battery. *European Journal of Work and Organizational Psychology, 15,* 241-261.

Kanning, U. P. & Schuler, H. (2014). Simulationsorientierte Verfahren der Personalauswahl. In H. Schuler & U. P. Kanning (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 216-256). Göttingen: Hogrefe.

Kasten, N. & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of situational judgment tests (SJTs). *European Journal of Psychological Assessment, 32,* 230-240.

Kirchgeorg, M. & Müller, J. (2013). Personalmarketing als Schlüssel zur Gewinnung, Bindung und Wiedergewinnung von Mitarbeitern. In R. Stock-Homburg (Hrsg.), *Handbuch strategisches Personalmanagement* (S. 73-90). Wiesbaden: Springer Gabler.

Klehe, U., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A. & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. *Human Performance, 25,* 273-302.

Kluger, A. M., Reilly, R. R. & Russel, C. J. (1991). Faking biodata tests: Are option-keyed instruments more resistant? *Journal of Applied Psychology, 76,* 889-896.

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H. & Hertel, G. (2015). How „situational" is judgment in situational judgment tests? *Journal of Applied Psychology, 100,* 399-416.

Landy, F. J. (2007). The validation of personnel decisions in the twenty-first century: Back to the future. In S. M. McPhail (Hrsg.), *Alternate validation strategies: Developing and leveraging existing validity evidence* (S. 409-426). San Francisco: Jossey-Bass.

Lievens, F. (2006). International situational judgment tests. In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests: Theory, measurement and application* (S. 279-300). Mahwah: Erlbaum.

Lievens, F. (2013). Adjusting medical school admission: Assessing interpersonal skills using situational judgment tests. *Medical Education, 47,* 182-189.

Lievens, F., Buyse, T. & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90,* 442-452.

Lievens, F., Chasteen, C. S., Day, E. A. & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91,* 247-258.

Lievens, F., De Corte, W. & Schollaert, E. (2008). A closer look at the frame-of-refenrence effect in personality scale scores and validity. *Journal of Applied Psychology, 93,* 268-279.

Lievens, F. & De Soete, B. (2012). Simulations. In N. Schmitt (Hrsg.), *The Oxford handbook of personnel assessment and selection* (S. 383-410). Oxford: Oxford University Press.

Lievens, F. & De Soete, B. (2015). Situational judgment tests. In N. J. Smelser & P. B. Baltes (Hrsg.), *International encyclopedia of the social & behavioral sciences* (S. 13-19). Oxford: Elsevier Science Ltd.

Lievens, F. & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology, 9,* 1-20.

Lievens, F., Peeters, H. & Schollaert, E. (2008). Situational judgment tests: a review of recent research. *Personnel Review, 37,* 426-441.

Lievens, F. & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91,* 1181-1188.

Lievens, F. & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology, 97,* 460-468.

Lievens, F., Sackett, P. R. & Buyse, T. (2009). The effect of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology, 94,* 1095-1101.

MacCann, C. & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion, 8,* 540-551.

MacCann, C., Ziegler, M. & Roberts, R. D. (2012). Faking in personality assessment: reflections and recommendations. In M. Ziegler, C. MacCann & R. D. Roberts, *New perspectives on faking in personality assessment* (S. 309-329). New York: Oxford University Press.

Marcus, B. (2011). *Personalpsychologie*. Wiesbaden: VS Verlag für Sozialwissenschaften.

McCarthy, J. M., van Iddekinge, C. H., Lievens, F., Kung, M., Sinar, E. F. & Campion, M. A. (2013). Do candidate reactions relate to job performance or affect criterion-related validities? A multistudy investigation of relations among reactions, selection test score, and job performance. *Journal of Applied Psychology, 98,* 701-719.

McCrae, R. R. & Costa, P. T. (2008). The five-factor theory of personality. In O. P. John, R. W. Robins & L. A. Pervin (Hrsg.), *Handbook of personality: Theory and research* (S. 159-181). New York: Guilford Press.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L. & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity. *Personnel Psychology, 60,* 63-91.

McDaniel, M. A., List, S. K. & Kepes, S. (2016). The „hot mess“ of situational judgment test construct validity and other issues. *Industrial and Organizational Psychology, 9,* 45-51.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A. & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730-740.

McDaniel, M. A. & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33,* 515-525.

McDaniel, M. A. & Whetzel, D. L. (2007). Situational judgment tests. In D. L. Whetzel & G. R. Wheaton (Hrsg.), *Applied measurement: Industrial psychology in human resources management* (S. 235-258). Mahwah: Erlbaum.

McDaniel, M. A., Whetzel, D. L., Hartman, N. S., Nguyen, N. T. & Grubb, W. L. (2006). Situational judgment tests: Validity and an integrative model. In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests: Theory, measurement and application* (S. 183-203). Mahwah: Erlbaum.

McDaniel, M. A., Whetzel, D. L. & Nguyen, N. T. (2006). Situational judgment tests in personnel selection. In L. Yan (Hrsg.), *Personnel Assessment Monograph* (S. 1-32). Alexandria: International Public Management Association for Human Ressources.

McFarland, L. A. (2013). Applicant reactions to personality tests: Why do applicants hate them? In N. D. Christiansen & R. P. Tett (Hrsg.), *Handbook of Personality at work* (S. 281-198). New York: Routledge.

McFarland, L. A. & Ryan, A. M. (2000). Variance in faking across non-cognitive measures. *Journal of Applied Psychology, 85,* 812-821.

McFarland, L. A., & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology, 36,* 979-1016.

Motowidlo, S. J. & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measures by situational judgment test. *Journal of Applied Psychology, 95,* 321-333.

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

Motowidlo, S. J., Hooper, A. C. & Jackson, H. L. (2006). A theoretical basis for situational judgment tests. In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests: Theory, measurement and application* (S. 57-81). Mahwah: Erlbaum.

Motowidlo, S. J. & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology, 66,* 337-344.

Mount, M. K., Witt, L. A. & Barrick, M. R. (2000). Incremental validity of empirically keyed biodata scales over GMA and the five factor personality constructs. *Personnel Psychology, 53,* 299-323.

Muck, P. M. (2013). Entwicklung von Situational Judgment Tests: Konzeptionelle Überlegungen und empirische Befunde. *Zeitschrift für Arbeits- und Organisationspsychologie, 57,*185-205.

Mueller-Hanson, R. A., Heggestad, E. D. & Thornton, G. C. (2006). Individual differences in impression management: An exploration of the psychological processes underlying faking. *Psychology Science, 3,* 288-312.

Mumford, M. D. (1999). Construct validity and backround data: Issues, abuses, and future directions. *Human Resource Management Review, 9,* 117-145.

Munshi, F., Lababidi, H. & Alyousef, S. (2015). Low- versus high-fidelity simulations in teaching and assessing clinical skills. *Journal of Taibah University Medical Sciences, 10,* 12-15.

Nguyen, N. T., Biderman, M. D. & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment, 13,* 250-260.

Nimon, K., Zientek, L. R. & Henson, R. K. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Psychology, 3,* 41-52.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory.* New York: McGraw-Hill.

Ones, D. S., Dilchert, S., Viswesvaran, C. & Judge, T. A. (2007). In support of personality assessment in oganizational settings. *Personnel Psychology, 60,* 995-1027.

Oostrom, J. K., Born, M. P., Serlie, A. W. & van der Molen, H. T. (2011). A multimedia situational test with a constructed-response format: Its relationship with personality, cognitive ability, job experience, and academic performance. *Journal of Personnel Psychology, 10,* 78-88.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5,* 343-355.

Ostendorf, F. & Angleitner, A. (2004). *NEO-Persönlichkkeitsinventar nach Costa und McCrae, Revidierte Fassung (NEO-PI-R).* Göttingen: Hogrefe.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J. & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89,* 187-207.

Ployhart, R. E. & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11,* 1-16.

Ployhart, R. E. & MacKenzie, W. I. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Hrsg.), *APA handbook of industrial and organizational psychology: Selecting and developing members for the organization* (S. 237-252). Washington: APA.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Application and data analysis methods.* Thousand Oaks: Sage.

Richman-Hirsch, W. L., Olson-Buchanan, J. B. & Drasgow, F. (2000). Examining the impact of administration medium on examinee perception and attitudes. *Journal of Applied Psychology, 85,* 880-887.

Robie, C., Schmit, M. J., Ryan, A. M. & Zickar, M. J. (2000). Effects of item context specifity on the measurement equivalence of a personality inventory. *Organizational Research Methods, 3,* 348-365.

Ryan, A. M. & Ployhart, E. (2014). A century of selection. *Annual Review of Psychology, 65,* 693-717.

Sackett, P. R. (2012). Faking in personality assessment: Where do we stand? In M. Ziegler, C. MacCann & E. Ployhart (Hrsg.), *New perspectives on faking in personality assessment* (S. 330-344). New York: Oxford University Press.

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

Schmidt, F. L., Hunter, J. E. & Urry, V. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473-485.

Schmit, M. J., Ryan, A. M., Stierwalt, S. L. & Powell, A. B. (1995). Frame of reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology, 80,* 607-620.

Schmit, M. J. & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78,* 966-974.

Schmitt, N. & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests: Theory, measurement and application* (S. 135-155). Mahwah: Erlbaum.

Schmitt, N. & Gilliland, S. W. (1992). Beyond differential prediction: Fairness in selection. In D. M. Saunders (Hrsg.), *New approaches to amployee management: Fairness in employee selection* (S. 21-46). Greenwich: JAI Press.

Schnabel, D., Kelava, A., Seifert, L. & Kuhlbrodt, B. (2014). Konstruktion und Validierung eines multimethodalen berufsbezogenen Tests zur Messung interkultureller Kompetenz. *Diagnostica, 61,* 3-21.

Schuler, H. (2009). Arbeits- und Organisationspsychologie: Berufseignungsdiagnostik und Personalauswahl. In G. Krampen (Hrsg.), *Psychologie - Experten als Zeitzeugen* (S. 180-194). Göttingen: Hogrefe.

Schuler, H. (2014). *Psychologische Personalauswahl: Eignungsdiagnostik für Personalentscheidungen und Berufsberatung* (4. Auflage). Göttingen: Hogrefe.

Shaffer, J. A. & Postlethwaite, B. E. (2012). A matter of context: A meta-analytical investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology, 65,* 445-494.

Snell, A. F., Sydell, E. J. & Lueke, S. B. (1999). Toward a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9,* 219-242.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D. & Lievens, F. (2016). Validity and reliability of situational judgment test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19,* 506-532.

Steinmayr, R., Schütz, A., Hertel, J. & Schröder-Abé, M. (2011). *MSCEIT: Mayer-Salovey-Caruso Test zur emotionalen Intelligenz*. Bern: Verlag Hans Huber.

Stemler, S. E. & Sternberg, R. J. (2006). Using situational judgment tests to measure practical intelligence. In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests: Theory, measurement and application* (S. 107-131). Mahwah: Erlbaum.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvarth, J. A., Wagner, R. K., Williams, W. M., Snook, S. A. & Grigorenko, E. L. (2000). *Practical intelligence in everyday life.* New York: Cambridge University Press.

Stevens, M. J. & Campion, M. A. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management, 25,* 207-228.

Tett, R. P. & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88,* 500-517.

Tett, R. P. & Guterman, H. A. (2000). Situational trait relevance, trait expression, and cross-situational consistency: Testing the principle of trait activation. *Journal of Research in Personality, 34,* 397-423.

Tett, R. P., Jackson, D. N. & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytical review. *Personnel Psychology, 44,* 703-742.

Tett, R. P., Simonet, D. V., Walser, B. & Brown, C. (2013). Trait activation theory: Applications, developments, and implications for person-workplace fit. In N. D. Christiansen & R. P. Tett (Hrsg.), *Handbook of personality at work* (S. 71-101). New York: Routledge.

Trippe, D. M. (2002). *An evaluation of the construct validity of situational judgment tests* (Master's thesis). Verfügbar unter: http://vtechworks.lib.vt.edu/

Tuzinski, K. (2013). Simulations for personnel selection: An introduction. In M. Fetzer & K. Tuzinski, *Simulations for personnel selection* (S. 1-16). New York: Springer.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58,* 6-20.

Vacha-Haase, T. & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development, 44,* 159-168.

Viswesvaran, C. & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59,* 197-210.

Wagner, R. K. & Sternberg, R. J. (1991). *Tacit knowledge inventory for managers: Users manual.* San Antonio: The Psychological Corporation.

Wang, L., MacCann, C., Zhuang, X., Liu, O. L. & Roberts, R. D. (2009). Assessing teamwork and collaboration in high school students. *Canadian Journal of School Performance, 24,* 108-124.

Weekley, J. A., Ployhart, R. E. & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests: Theory, measurement and application* (S. 157-182). Mahwah: Erlbaum.

Weekley, J. A. & Ployhart, R. E. (2006). An introduction to situational judgment testing In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests: Theory, measurement and application* (S. 1-10). Mahwah: Erlbaum.

Weuster, A. (2008). *Personalauswahl: Anforderungsprofil, Bewerbersuche, Vorauswahl und Vorstellungsgespräch.* Wiesbaden: Gabler.

Whetzel, D. L. & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19,* 188-202.

Wilkinson, L. & American Psychological Association (APA) Task Force on Statistical Inference (1999). Statistical methods in psychological journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Zickar, M. J., Gibby, R. E. & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7,* 168-190.

Ziegler, M. & Kemper, C. J. (2013). Extreme response styles and faking: Two sides of the same coin. In P. Winkler, N. Menold & R. Prost (Hrsg.), *Interviewers' deviations in surveys: Impact, detection and prevention* (S. 217-233). Frankfurt: PL Academic Research.

Zimmerman, D. W., Zumbo, B. D. & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53,* 33-49.

Anhang

- **Anhang A:** Korrelationsdiagramme der Antwortoptionen
- **Anhang B:** Eigenständigkeitserklärung

Anhang A

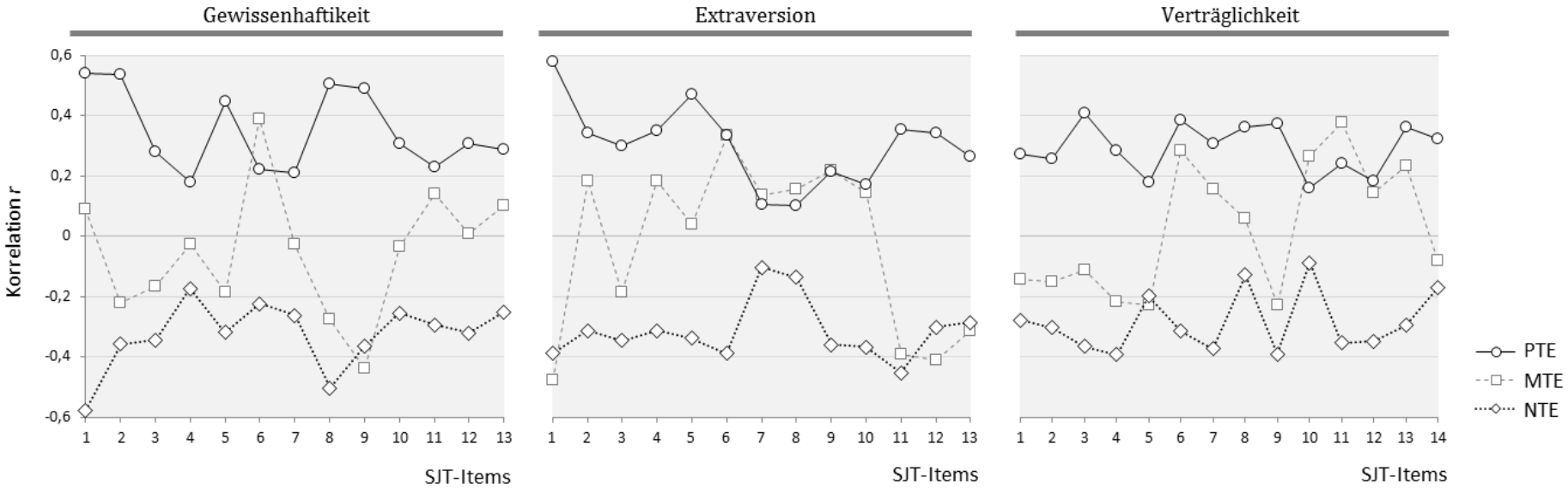

Abbildung 5. Korrelationen der Antwortoptionen der SJT Items mit der jeweils entsprechenden Skala des NEO-PI-R. PTE = *positive trait expression*, MTE = *moderate trait expression*, NTE = *negative trait expression*. Berechnungen anhand des Datensatzes aus Artikel 2 ($N = 162$)

Anhang B

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Bei der Auswahl und Auswertung folgenden Materials haben mir die nachstehend aufgeführten Personen in der jeweils beschriebenen Weise entgeltlich/unentgeltlich geholfen.

1. Zur Bestimmung der Interrater-Reliabilität im Rahmen der Reliabilitätsgeneralisierung (Artikel 1) kodierte Robert Klimanek (zum damaligen Zeitpunkt studentische Hilfskraft an der Universität Osnabrück) die Studien- und Testcharakteristika der einbezogenen Studien.

2. Die Datensätze wurden zu einem großen Teil im Rahmen von studentischen Abschlussarbeiten erhoben, die von mir betreut wurden. An der Datenerhebung beteiligt waren Dorina Gottschlich, Svenja Heuermann, Viktoria Henke, Natalia Krüger, und Hanna Schlautmann.

3. Zusätzlich an der Datenerhebung beteiligt waren Verena Becker und David Peukert (beide im Rahmen eines Praktikums an der Albert-Ludwigs-Universität Freiburg).

Weitere Personen waren an der inhaltlichen materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- und Beratungsdiensten in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Freiburg, 2017