Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

Online Self-Assessment in der Personalauswahl

Entwicklung eines Instruments zur Selbstselektion mittels 360° Videos

MASTER-ARBEIT

2022 / 2023

Autorin
Oberholzer, Kim Nina

Begleitperson

Dr. Christ, Oliver

Praxispartnerin
Gelateria di Berna AG
Kontaktperson Käser, Andreas

Abstract

In this master thesis, an online self-assessment for applicants was developed to support Gelateria di Berna in the annual recruitment of their sales staff. In a sequential mixed-methods study design, a Situational Judgment Test was developed using Flanagan's Critical Incident Technique (1956), filmed with a 360° camera and enriched with game elements. In an online survey, the instrument was tested in a real job application setting in order to assess the validity of individual predictors by means of linear regressions with regard to certain primary factors of the Social Skills Inventory (Inventar Sozialer Kompetenzen, ISK) by Kanning (2009). It was intended to contribute to the generally prevailing research deficit in the area of the validity of digital solutions in personnel selection. The positive evaluations of the online self-assessment by the respondents confirms the high acceptance of such digital formats in the recruitment process. However, no significant effects could be found for variance clarification through the developed instrument.

Keywords: Online Self-Assessment, Situational Judgment Test, personnel selection, social competencies

Zusammenfassung

Im Rahmen der vorliegenden Masterarbeit wurde ein Online Self-Assessment für Stelleninteressierte entwickelt, um die Gelateria di Berna bei der jährlichen Rekrutierung von Verkaufsmitarbeitenden zu unterstützen. In einem sequenziellen Mixed-Methods-Studiendesign wurde unter Anwendung der Critical Incident Technique von Flanagan (1956) ein Situational Judgment Test entwickelt, mit einer 360°-Kamera verfilmt und mit Game-Elementen angereichert. In einer Online Umfrage wurde das Instrument im realen Bewerbungssetting getestet, um die Validität einzelner Prädiktoren mittels linearer Regressionen bezüglich bestimmter Primärfaktoren des Inventars sozialer Kompetenzen (ISK) von Kanning (2009) zu bewerten. Damit sollte ein Beitrag zum allgemein vorherrschenden Forschungsdefizit im Bereich der Validität von digitalen Lösungen in der Personalauswahl geleistet werden. Die positiven Bewertungen des Online Self-Assessments durch die Befragten bestätigten die hohe Akzeptanz gegenüber solchen digitalen Formaten im Personalbeschaffungsprozess. Es konnten jedoch keine signifikanten Effekte zur Varianzaufklärung durch das entwickelte Instrument gefunden werden.

Schlüsselbegriffe: Online Self-Assessment, Situational Judgment Test, Personalauswahl, soziale Kompetenzen

Anzahl Zeichen (inkl. Leerzeichen, exkl. Anhang): 212'245

Inhaltsverzeichnis

1 Einleitung

Nach zwei Jahren Covid-Pandemie profitieren aktuell viele Berufsgruppen von der wirtschaftlichen Erholung. Laut einer Medienmitteilung des Bundesamt für Statistik vom Mai 2022 (BFS, 2022) bewegte sich die Anzahl offener Stellen in der Schweiz im ersten Quartal des Jahres 2022 im Rekordbereich. Der Anstieg im Vergleich zum entsprechenden Vorjahresquartal betrug +60.4%. Dies entspricht 100'000 offener Stellen, was es so in der Schweiz noch nicht gab. Während die Schweizerische Gesamtbeschäftigung um 2.5% gestiegen ist, haben die Schwierigkeiten bei der Personalrekrutierung von gelernten Arbeitskräften zugenommen. 37.5% der befragten Unternehmen haben angegeben, dass sie nur schwer oder gar kein qualifiziertes Personal gefunden haben. Besonders betroffen sind laut der Studie das Gesundheitswesen, Informatikberufe sowie das Gastgewerbe (BFS, 2022). Gemäss einer Studie von Ryf, Siegenthaler, Fasnacht und Fichter (2022) wird dieser Fachkräftemangel als längerfristige Herausforderung angesehen. Es ist deshalb nicht erstaunlich, dass das Forschungsinteresse im Bereich der Rekrutierung in diesen Zeiten hoch ist. Gerade im Zusammenhang mit der fortschreitenden Digitalisierung ergeben sich neue, vorteilhafte Möglichkeiten, sich mit innovativen und modernen Rekrutierungsstrategien von anderen Unternehmen abzuheben. Überraschenderweise gibt es aber sehr wenig Forschung zu diesem Thema im Bereich des Gastgewerbes. Überraschend deshalb, weil gerade dort die Notwendigkeit für eine erfolgreiche und kosteneffiziente Rekrutierung hoch ist. Denn hohe Fluktuation und saisonale Schwankungen sorgen für einen hohen Personalbedarf in dieser Branche (Ladkin & Buhalis, 2016).

1.1 Praxispartnerin und Ausgangslage

Mit den eben geschilderten Herausforderungen sieht sich auch die Praxispartnerin der vorliegenden Masterarbeit konfrontiert. Die Gelateria di Berna AG (im Folgenden mit GdB abgekürzt) ist eine Firma im Bereich der Lebensmittelproduktion. In den insgesamt acht Filialen in der Schweiz werden Gelati frisch produziert und direkt an die Endkundschaft

verkauft. Während der Hauptsaison zwischen März und Oktober werden dazu über 200 Verkaufsmitarbeitende beschäftigt. Sie sind die ersten Ansprechpersonen in den Filialen und das Aushängeschild der Firma. Durch die starke Saisonalität des Gelato-Geschäfts können aber nur wenige Mitarbeitende ganzjährig in einer Festanstellung angestellt werden. Alle anderen arbeiten in befristeten Arbeitsverhältnissen. Zwar können rund ein Drittel der ehemaligen Verkaufsmitarbeitenden jeweils für die anschliessende Saison wiedergewonnen werden, dennoch müssen jährlich über 200 Bewerbende in Einstellungsgesprächen kennengelernt und von der Stelle im Verkauf sowie der GdB als Arbeitgeberin überzeugt werden. Da die Stelle im Verkauf während den Sommermonaten viele junge Studierende anspricht, die neben dem Studium und in den Semesterferien arbeiten möchten, sind unter den Bewerbungen viele junge Personen, für die es die erste Anstellung überhaupt ist. Auch deshalb muss während den Einstellungsgesprächen ein grosser Informationsaufwand betrieben werden, um falsche Erwartungen vorzubeugen und passende Bewerbende für die Verkaufsstellen zu gewinnen. Erschwerend kommt dazu, dass es viele vergleichbare Stellenangebote auf den Stellenmarkt gibt und die Stelleninteressierten zwischen verschiedenen Möglichkeiten frei auswählen können. Es ist deshalb nicht selten, dass nach bereits erfolgtem, zeitintensivem Ressourceneinsatz durch die GdB einige Bewerbende ihre Bewerbung aufgrund eines anderen Jobangebots zurückziehen.

In dieser deutlichen Verschiebung von einem Arbeitgebermarkt zu einem Arbeitnehmermarkt ist es heute für Arbeitgebende umso wichtiger zu verstehen, was sich Bewerbende von Unternehmen wünschen, um die besten Talente für sich zu gewinnen. Mit der Digitalisierung ist der Personalbeschaffungsprozess von Unternehmen zu Unternehmen nicht nur individueller geworden, sondern auch asynchroner, schneller, diffuser und weniger linear (Verhoeven, 2020). Unter anderem werden in diesem Zusammenhang schon häufig sogenannte Online Assessments verwendet, mit dem Zweck, die Personalauswahl effizienter und valider zu gestalten. Noch relativ neu in der Personalauswahl sind hingegen sogenannte Online Self-Assessments und Recrutainment-Methoden. Diese Formate

bestechen nicht nur durch ihre testende Funktionsweise, sondern verfügen auch über eine informierende Komponente (Ott, Ulfert & Kersting, 2017). Ausserdem sollen sie die Neugierde von potenziellen Stelleninteressierten wecken und verfügen je nach Format über eine «Anlockfunktion» (Diercks, 2021). Gleichzeitig sollen durch die Nutzung solcher Selbstselektionstools Vorurteile oder überzogene Erwartungen relativiert werden (Hiltmann, 2013).

Während sich die Anwendung digitaler Lösungen in der Personalauswahl sehr rasch verbreitet und etabliert, hinkt laut Review von Woods, Ahmed, Nikolaou, Costa und Anderson (2020) die Forschung zu deren Validität noch deutlich hinterher. Da der Einsatz solcher Verfahren in der Personalauswahl meist das Ziel verfolgt, die spätere Arbeitsleistung (Job Performance) von Bewerbenden einzuschätzen, werden Online Assessments häufig auf Grundlage eines Situational Judgment Tests (SJT) entwickelt. Als simulationsorientiertes Verfahren werden dabei typische Situationen aus dem Berufsalltag rekonstruiert und den Bewerbenden in Form eines Textes oder Videos präsentiert. Zwar wird die Realitätsnähe speziell von textbasierten SJT im Vergleich mit anderen simulationsorientierten Verfahren als eher gering eingeschätzt, dennoch geniesst das Verfahren unter anderem wegen des verhältnismässig ökonomischen Auswertungsaufwand grosse Beliebtheit (Weekley, Hawkes, Guenole & Ployhart, 2015). Studien zur prädiktiven Validität von SJT finden sich überwiegend im medizinischen Kontext (z.B. Lievens & Sackett, 2012; Webster, Paton, Crampton & Tiffin, 2020). Aber auch im Ausbildungssektor (Bardach, Rushby, Kim, & Klassen, 2021) oder in der Sicherheitsbranche (Leeds, Griffith, & Frei, 2003) kommen SJT schon erfolgreich in der Personalauswahl zum Einsatz.

Mit dem Hintergrund des prognostizierten, anhaltenden Fachkräftemangels in einer Branche, die trotz hoher Fluktuation und starker Saisonalität bisher nur wenig Forschungsinteresse geniesst, soll sich die vorliegende Masterarbeit die aktuellen Erkenntnisse aus der Forschung zur Hilfe nehmen und die GdB einen Schritt weiter in Richtung einer erfolgreichen und kosteneffizienten Rekrutierung bringen. Es soll ein Online

Self-Assessment entwickelt werden, das zum einen die Neugierde von Stelleninteressierten weckt und gleichzeitig mit psychometrischer Fundiertheit dazu beiträgt, dass sich in einem hart umkämpften Arbeitnehmermarkt möglichst viele passende Stelleninteressierte mittels Selbstselektion zu einer Bewerbung für die Stelle als Verkaufsmitarbeitende bei der GdB entscheiden. Durch eine erste Testung des zu entwickelnden Instruments im realen Bewerbungssetting soll zudem ein Beitrag zum vorherrschenden Forschungsdefizit im Bereich der Validität digitaler Verfahren in der Personalauswahl geleistet werden.

1.2 Aufbau der Arbeit

Nachdem in Kapitel 1 die Ausgangslage und Problemstellung der Masterarbeit geschildert wurden, widmet sich Kapitel 2 dem theoretischen Hintergrund. Es wird zuerst ein Überblick über die Personalauswahl in Zeiten der Digitalisierung gegeben, bevor dann näher auf Online Assessments und den Situational Judgment Test eingegangen wird. Die theoretischen Grundlagen werden mit empirischen Studienergebnissen untermauert. Kapitel 2 schliesst mit der begründeten Zielsetzung der Masterarbeit und den daraus abgeleiteten Leitfragen. In Kapitel 3 werden das gewählte Untersuchungsdesign, das methodische Vorgehen und die eingesetzten Instrumente beschrieben und begründet. Die Ergebnisse daraus werden zuerst in Kapitel 4 präsentiert und anschliessend in Kapitel 5 interpretiert sowie in Bezug auf den theoretisch-empirischen Hintergrund diskutiert. Es wird ebenfalls auf Limitationen der vorliegenden Masterarbeit hingewiesen und ein Ausblick auf weiterführende Forschungsfelder gegeben. Ein Fazit rundet die Arbeit ab.

2 Theoretischer Hintergrund

Das Kapitel 2 setzt den theoretischen und empirischen Bezugsrahmen der vorliegenden Masterarbeit. Es soll schrittweise an die Zielsetzung heranführen und die relevanten und aktuellen Konzepte sowie Forschungsergebnisse vorstellen. Nachdem zuerst mithilfe des Personalbeschaffungsprozess ein grober Überblick über die zentralen Begrifflichkeiten in der Personalauswahl im digitalen Zeitalter gegeben wird, soll im Anschluss das Online Assessment und dessen unterschiedliche Gestaltungsformen vorgestellt werden. Der Situational Judgment Test wird als zugrundeliegendes, psychometrisches Verfahren beschrieben, bevor zum Abschluss des Kapitels die Zielsetzung der Masterarbeit und die Leitfragen aus der Theorie abgeleitet werden.

2.1 Personalauswahl in Zeiten der Digitalisierung

Die Rekrutierung und Personalauswahl ist für Unternehmen schon seit Beginn ein bedeutender und vor allem teurer Faktor in der Geschäftsführung. Die Digitalisierung hatte dabei in den letzten Jahren einen grossen Einfluss darauf, wie Personalverantwortliche arbeiten und Stelleninteressierte nach einem Job suchen (Nikolaou, 2021). In den folgenden Abschnitten wird ein Überblick über die aktuellen Entwicklungen und Trends im Personalbeschaffungsprozess gegeben und der Begriff *Candidate Experience* als wichtiger Bestandteil im modernen Personalbeschaffungsprozess vorgestellt.

2.1.1 Personalbeschaffungsprozess

Im Zeitalter der unterschiedlichsten Job- und beruflichen Netzwerkplattformen ist es für Stellensuchende so einfach wie nie, interessante Stellenanzeigen und Organisationen zu finden, zu vergleichen und sich mit wenigen Klicks zu bewerben. Doch mit der steigenden Anzahl Bewerbungen pro zu besetzender Stelle steigt auch die Anzahl an unqualifizierten Bewerbungen (Black & Esch, 2020). Zur effizienteren und valideren Personalauswahl greifen Unternehmen deshalb immer häufiger auf digitale Instrumente und Tools zurück. Dabei geht es aber weniger darum, den Menschen aus dem Prozess zu

verdrängen, sondern es soll vielmehr die bisherige Diagnostik mit digitalen Tools unterstützt werden (Petschar & Zavrel, 2016). Während sich die Anwendung digitaler Formate in der Personalauswahl sehr rasch verbreitet und etabliert, hinkt aber laut Review von Woods et al. (2020) die Forschung zur Validität solcher Tools noch deutlich hinterher. So wird fälschlicherweise automatisch davon ausgegangen, dass die Funktionsweise und Validität von traditionellen Methoden, wie beispielsweise einem konventionellen Interview, einfach auf digitale Formate, in diesem Fall ein digitales Interview, übertragbar sind. Laut der Autorenschaft müssen sich digitale Formate aber unabhängig von traditionellen Formaten betreffend ihrer psychometrischen Effektivität erst einmal beweisen. Erst wenn evidenzbasiert gezeigt werden kann, dass digitale Formate im Kontext der Personalauswahl die gewünschten Validitäten erreichen oder bestehende Methoden sogar übertreffen, kann der Wechsel zu digitalen Alternativen gerechtfertigt werden (Woods et al., 2020).

Digitale Methoden können zu ganz unterschiedlichen Zeitpunkten während der Personalauswahl zum Einsatz kommen. Je nach Autor und Fokus werden in der Literatur für den Prozess der Personalbeschaffung sowie einzelne Phasen daraus unterschiedliche Begriffe synonym verwendet. Klassischerweise werden aber laut Nikolaou (2021) vier Phasen unterschieden: Attraction, Screening, Selection und Onboarding (siehe Abbildung 1). Jede der vier Phasen hat dabei eine spezifische Funktion, die mithilfe von digitalen Elementen unterstützt werden kann.

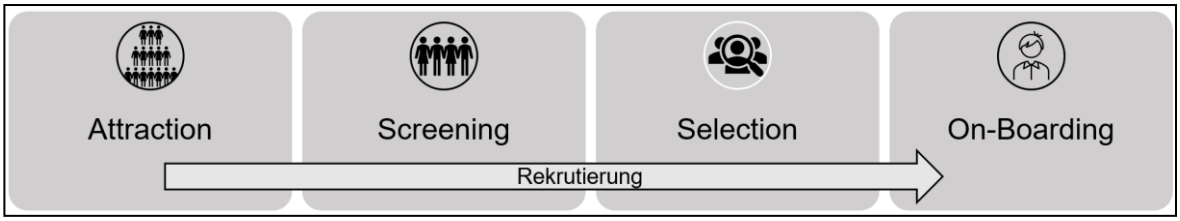

Abbildung 1. Phasen des Personalbeschaffungsprozesses (Nikolaou, 2021; eigene Darstellung)

Die Attraction-Phase (engl. Anziehung) umfasst dabei in der kurzen Frist alle Massnahmen, die gezielt möglichst viele und qualitativ hochwertige Bewerbungen für bestimmte offene Stellen generieren sollen (Chapman & Mayers, 2015). Neben eher klassischen Stelleninseraten auf den Karriereseiten der Unternehmen, werden auch Social

Networking Webseiten und Social Media genutzt, um offene Jobangebote zu bewerben (Nikolaou, 2014). Über standardisierte Online Bewerbungsformulare können sich Stelleninteressierte dann direkt mit wenigen Klicks bewerben. Je einfacher und sicherer (bezüglich Datenschutz) solche Formulare von den Stelleninteressierten wahrgenommen werden, desto eher bewerben sie sich (Bauer, Truxillo, Tucker, Weathers, Bertolino, Erdogan und Campion, 2006). Da Online Bewerbungsformulare meist auch der erste richtige Kontaktpunkt von Bewerbenden mit einer Unternehmung sind, haben sie grossen Einfluss darauf, welchen ersten Eindruck die Bewerbenden von der Unternehmung gewinnen (Truxillo, Bauer, McCarthy, Anderson & Ahmed, 2018). Ein guter erster Eindruck ist deshalb besonders wichtig, da Bewerbende mit dem Absenden ihrer Unterlagen diejenigen sind, die die erste Entscheidung während des Personalbeschaffungsprozess treffen. Um Stelleninteressierte bei ihrer Entscheidung bestmöglich zu unterstützen, stellen Unternehmen immer häufiger auch digitale Tools zur Verfügung, die die Selbstselektion der potenziellen Kandidatinnen und Kandidaten unterstützen sollen. Neben einfacheren Entscheidungshilfen wie den sogenannten *Matching-Tools* (Ott et al., 2017) gibt es innerhalb der Attraction-Phase auch Formate, die auf spielerische Art und Weise die Neugierde von Stelleninteressierten wecken sollen (Larson, 2019). Dazu folgt in Kapitel 2.2 mehr.

Nachdem erste Informationen zu den Bewerbenden bei einer Unternehmung eingegangen sind, beginnt die Screening-Phase (engl. Durchleuchtung). Dabei kommt es zu einer ersten Vorauswahl seitens der potenziellen Arbeitgebendem innerhalb der eingegangenen Dossiers, um qualifizierte von unqualifizierten Bewerbungen zu unterscheiden. Dies geschieht unter anderem aufgrund der abgefragten Informationen im Online Bewerbungsformular. Durch eine standardisierte Abfrage wird der Screening-Prozess nicht nur effizienter und dadurch kostengünstiger, sondern auch objektiver (Konradt, Warszta & Ellwart, 2013). Im digitalen Setting wird aber immer häufiger auch auf Informationen von den Social Media Profilen der Bewerbenden zurückgegriffen. Die

Berücksichtigung des digitalen Fussabdrucks der Bewerbenden im Bewerbungsprozess wird unter dem Begriff Cybervetting zusammengefasst und stösst bei Bewerbenden vor allem beim Zurückgreifen auf Informationen von nicht-beruflichen sozialen Netzwerken wie beispielsweise Facebook auf wenig Akzeptanz (Cook, Jones-Chick, Roulin & O'Rourke, 2020).

Unter der Selection-Phase (engl. Auswahl) werden sämtliche Methoden zusammengefasst, die dabei helfen sollen, die nach dem Screening als passend befundenen Bewerbenden in den für die zu besetzende Stelle relevanten, individuellen Persönlichkeitseigenschaften, kognitiven Fähigkeiten sowie jobspezifischen Kompetenzen zu unterscheiden. Was vor der Digitalisierung mittels Paper-Pencil-Tests (schriftlichen Tests) abgefragt wurde, wird heute häufig in ein webbasiertes Format, ein sogenanntes Online Assessment übersetzt. Wie bereits eingangs erwähnt, gibt es zur Austauschbarkeit von schriftlichen und digitalen Formaten noch nicht ausreichend Evidenz (Woods et al., 2020). Le Corff, Gingras und Busque-Carrier (2017) konnten in ihrer Studie aber beispielsweise zeigen, dass die Ergebnisse von internetbasierten Testungen der Big Five Persönlichkeitseigenschaften eine sehr hohe Übereinstimmung mit dem Paper-Pencil-Format aufweisen. Auch für klassische Jobinterviews werden digitale Alternativen geschaffen. Anstelle oder zusätzlich zu Online Assessments werden immer häufiger digitale Interviews durchgeführt. Neben Interviews über Videokonferenzen sind auch asynchrone Interviews immer häufiger Teil digitaler Personalbeschaffungsprozesse. Dabei werden die Antworten der Bewerbenden aufgezeichnet und anschliessend durch ausgewählte Raterinnen und Rater oder mittels künstlicher Intelligenz automatisch ausgewertet. Erste Studien zur Validität dieses digitalen Formats konnten zeigen, dass asynchrone Interviews zu besseren Ratings führen. Es wird vermutet, dass dieses Resultat mit der längeren Vorbereitungszeit bei asynchronen Interviews verglichen zum digitalen Live-Format zu erklären ist (Langer, König und Krause, 2017). Aus Sicht der Bewerbenden zeigen mehrere Studien, dass digitale Interviews als weniger fair und weniger vorteilhaft bewertet werden,

da beispielsweise die Körpersprache, aber auch verbale und nonverbale Hinweise verloren gehen (Woods et al., 2020).

Die letzte Phase des Personalbeschaffungsprozess ist das On-Boarding. Dieses beginnt ab dem ersten Tag der neuen Mitarbeitenden in einem Unternehmen. Während beispielsweise On-the-Job-Trainings und Mentorings bereits als signifikante Faktoren für eine erfolgreiche Sozialisation von neuen Mitarbeitenden erkannt wurden, wird auch in diesem Bereich immer öfter auf digitale Alternativen wie Online-Training, E-Mentoring, Intranet bis hin zu Social Media Seiten der betreffenden Unternehmen gesetzt (Nikolaou, 2021). Mit der fortschreitenden Technologie wird gerade im Bereich von Schulungen intensiv an der Wirksamkeit von Virtual Reality (VR) Formaten geforscht (Hamilton, McKechnie, Edgerton & Wilson, 2020). Dazu folgt im weiteren Verlauf der Arbeit noch ein eigener Abschnitt (siehe Kapitel 2.3.2).

Über alle vier Phasen des Personalbeschaffungsprozess hinweg geht es langfristig darum, ein attraktives Arbeitgeberimage aufzubauen und zu pflegen, um dadurch indirekt qualifizierte Personen auf das Unternehmen aufmerksam zu machen und auf dem Arbeitsmarkt bestehen zu können. Für diesen langfristigen Ansatz wird in der Literatur häufig der übergeordnete Begriff *Rekrutierung* verwendet (Nikolaou, 2021).

2.1.2 Candidate Experience

Während Unternehmen früher noch in der Position waren, sich in der Screening-Phase aus einer grossen Anzahl Bewerbungen diejenigen aussuchen zu können, die am besten zur ausgeschriebenen Stelle passten, so wird die Situation heute etwas salopp mit «Post and Pray» (engl. für Ausschreiben und Beten) umschrieben (Diercks, 2021). Mit dem zunehmenden Wettbewerb um qualifizierte Bewerbende entsteht für Unternehmen die Notwendigkeit, ihre vakanten Stellen als Produkt zu verstehen und auf dem Markt möglichst erfolgreich zu positionieren (Kanning, 2009). Gleichzeitig sind die Akzeptanz der Verfahren sowie bewerberfreundliche und wertschätzende Prozesse wichtiger denn je zuvor (Fellner, 2019; Petschar & Zavrel, 2016). Seit einigen Jahren wird der Begriff *Candidate Experience*

als integraler Bestandteil des Personalbeschaffungsprozess verstanden. Als Candidate Experience wird laut Verhoeven (2016) der Gesamteindruck bezeichnet, den potenzielle Bewerbende im Rahmen des gesamten Personalbeschaffungsprozesses und darüber hinaus vom potenziellen Arbeitgebenden erhalten. Dabei geht es um das individuelle Erleben der Bewerbenden an allen direkten und indirekten Kontaktpunkten mit dem Unternehmen. Dass diese Bewertung durch die Bewerbenden nicht nur bei der ersten Entscheidung dazu, sich überhaupt zu bewerben, in der Attraction-Phase relevant für jedes Unternehmen ist, zeigen folgende Zahlen: 78% der positiven und 59% der negativen Erfahrungen im Bewerbungsprozess werden mit engen Freunden und Arbeitskollegen geteilt. 51% der positiven und 31% der negativen Erfahrungen werden sogar über das Internet mit der Öffentlichkeit geteilt (Talent Board, 2021).

Ein erster möglicher Anknüpfungspunkt um die Candidate Experience im digitalen Personalbeschaffungsprozess zu optimieren, ist die Gestaltung des Wegs bis hin zum Abschicken der Bewerbung, also die Attraction-Phase. Petschar und Zavrel (2016) vergleichen diesen Prozess mit dem E-Commerce. Sind potenzielle Käuferinnen und Käufer erst einmal im Webshop (oder Stelleninteressente auf der Karriereseite) angekommen, so sollten sie nicht erst lange Formulare ausfüllen müssen, bevor sie ihren Kauf (ihre Bewerbung) abschliessen können. Ebenfalls sollten sie alle relevanten Informationen auf der Webseite nicht lange suchen müssen, sondern idealerweise automatisch vorgeschlagen bekommen. In der Screening- und Selection-Phase erwarten Bewerbende laut Brenner (2016) vor allem einen ausreichenden Stellenbezug bei der Formulierung von Fragen und Testinhalten sowie die Chance zu zeigen, was sie können (chance to perform). Während ersteres unabhängig von der Anwendung von Technologie zu beachten ist, ist gerade beim Einsatz von digitalen Auswahlverfahren besonders auf die wahrgenommene Nutzerfreundlichkeit (Usability) zu achten. Wenn für die Anwendung des Tools überdurchschnittliche IT-Kenntnisse nötig sind, so leidet nicht nur die wahrgenommene Nutzerfreundlichkeit und Fairness aus der Perspektive von Bewerbenden, sondern es hat

auch Einfluss auf die Resultate (Ott et al., 2017). Brenner (2016) verweist zudem auf den wahrgenommenen Nutzen von eingesetzten Tools im Auswahlprozess. So kann ein asynchrones Interview beispielsweise den Vorteil mitbringen, dass keine Terminabsprachen im Vorfeld nötig sind und das Interview zeit- und ortflexibel durchgeführt werden kann. Und nicht zuletzt hat auch das Feedback oder die automatisierte Rückmeldung eines Online-Assessments einen Einfluss auf die Akzeptanz der Bewerbenden (Ott et al., 2017).

Grundsätzlich ist der Personalbeschaffungsprozess und damit die Candidate Experience heute nicht nur digitaler, sondern dadurch auch asynchroner (beispielsweise durch das asynchrone Interview), schneller (Geschwindigkeit als Wettbewerbsvorteil bei der Rekrutierung), weniger aufwendig (Bewerbung durch einen Klick) (Verhoeven, 2020) und damit auch kostengünstiger (Okolie & Irabor, 2017). Durch den Einsatz von Online Assessments oder Recrutainment-Methoden (dazu später mehr) wird der Prozess ausserdem individueller von Unternehmen zu Unternehmen, diffuser und weniger linear. Es gibt keinen klassischen Bewerbungsprozess mehr, sondern die einzelnen Phasen überschneiden sich oder die Abfolge wird leicht abgeändert. Ein ausreichendes Verständnis davon, was sich Bewerbende von Unternehmen wünschen und die Digitalisierung dementsprechend einzusetzen ist deshalb unabdingbar, um die besten Talente für sich zu gewinnen (Verhoeven, 2020).

2.2 Online Assessments im Personalbeschaffungsprozess

Da Online Assessments in allen Phasen des Personalbeschaffungsprozesses zum Einsatz kommen und zu einer effizienteren und valideren Personalauswahl beitragen können, wird im digitalen Zeitalter sehr häufig und gerne auf dieses Format zurückgegriffen. Der Begriff *Online Assessment* (engl. Einschätzung oder Bewertung am Laptop oder mobilen Endgerät) meint in diesem Kontext grundsätzlich jede online durchgeführte Diagnostik, welche die berufliche Eignung von Bewerbenden feststellen soll (Konradt & Sarges, 2003). Dazu gehören beispielsweise klassische Online Tests und Fragebögen,

Online Case Studies, gamifizierte Verfahren oder auch der Einsatz von Avataren sowie Stimmanalysen (Fellner, 2019). Neben einer hohen Auswertungs- und Interpretationsobjektivität bieten Online Assessments je nach Anwendungsgebiet Vorteile wie zeitliche und räumliche Unabhängigkeit, Anonymität, ein modernes Image des Anbieters sowie ökonomische Vorteile (Ott et al., 2017). In Rahmen der vorliegenden Masterarbeit soll der Begriff Online Assessment, wie von Ott et al. (2017) vorgeschlagen, als Oberbegriff verwendet werden, der unterschiedliche Einsatzziele und Gestaltungsformen zulässt. In den folgenden Abschnitten wird eine Übersicht über verschiedene Differenzierungselemente von Online Assessments gegeben.

2.2.1 Einsatzziele und Gestaltungselemente

Online Assessments können in allen Phasen des Personalbeschaffungsprozesses zum Einsatz kommen und alle Anwendungsformen funktionieren grundsätzlich nach dem gleichen Prinzip: Zu Beginn steht das Interesse einer Unternehmung an einer bewerbenden Person oder das Interesse einer Person an einer Tätigkeit in einer Unternehmung. Mithilfe eines entsprechenden Online Assessments gelangt nun die Unternehmung an die benötigten Informationen über die Person respektive der oder die stelleninteressierte Person an die Anforderungen der betreffenden Tätigkeit. Aufgrund dieses Informationsgewinns kommt es dann zu einer Entscheidung (Ott et al., 2017). Bei Online Assessments handelt es sich um Instrumente der sogenannten Negativselektion. Das heisst, das primäre Ziel ist nicht die Identifikation der Bewerbenden, die eingestellt werden, sondern zunächst sollen diejenigen gefunden oder «abgeschreckt» werden, die mit grosser Wahrscheinlichkeit im weiteren Verlauf des Auswahlprozess als nicht geeignet bewertet werden. Die finale Entscheidung, ob Bewerbende im Auswahlprozess bleiben oder nicht, wird aber aus ethischen und diagnostischen Gründen normalerweise nicht durch einen dahinterliegenden Algorithmus getroffen. Vielmehr sollen die Ergebnisse als zusätzliche Beurteilungsperspektive im ganzheitlichen Auswahlprozess verwendet werden (Verhoeven, 2020).

In Abbildung 2 findet sich eine Gegenüberstellung der möglichen Einsatzziele und Gestaltungselemente von Online Assessments. Während es in der Screening- und Selection-Phase darum geht, als Unternehmen die Bewerbenden zu testen (Fremdselektion), um Unterschiede in den für die betreffende Stelle relevanten Bereichen zu finden, sollen sich Stelleninteressierte bei Self-Assessments nicht nur selbständig testen, sondern sich auch über die betreffende Stelle und das Unternehmen informieren können. Durch die Interaktivität mit den Inhalten im Online Self-Assessment werden die gezeigten Informationen anders zur Kenntnis genommen als durch Auflistungen in Broschüren oder Webseiteneinträgen, was zu einer tieferen Auseinandersetzung mit den Anforderungen und Voraussetzungen führt (Ott et al., 2017). Ein weiteres Unterscheidungsmerkmal von Online (Self-)Assessments ist die Art der Darbietungsweise. Die Informationen, die im Screening und bei Online Self-Assessments verarbeitet werden, werden meist von zu Hause unbeaufsichtigt bearbeitet. In der Selection-Phase müssen die Ergebnisse aber eindeutig einer Person zugewiesen werden können, weshalb die Testungen häufig unter Aufsicht vorgenommen werden. Was das genau heisst, wird in Kapitel 2.2.4 noch genauer beschrieben.

	Gestaltung					
	Selektion		Funktionsmechanismus		Darbietungsweise	
Ziel des Online-Assessments	Selbst	Fremd	Testen	Informieren	Mit Aufsicht	Ohne Aufsicht
Personal-Vorauswahl (Screening)		●	●			●
Personal-Endauswahl (Selection)		●	●		●	
Self- Assessment	●		●	●		●

Abbildung 2. Einsatzziele und Gestaltungselemente von Online Assessments nach Ott et al. (2017)

2.2.2 Fremdselektion

Werden Online Assessments von Unternehmen in der Screening- oder Selection-Phase zur Unterscheidung der Bewerbenden aufgrund bestimmter Informationen genutzt, spricht man von Fremdselektion. Die internetbasierten, diagnostischen Verfahren sind dabei

in der Regel nicht frei zugänglich, sondern werden den Bewerbenden während dem laufenden Auswahlprozess zur Durchführung freigegeben (Diercks & Kupka, 2013). Der Schwerpunkt der überprüften Merkmale liegt oftmals auf der sogenannten kognitiven Leistungsfähigkeit. Diese kann ganz unterschiedlich getestet werden, beispielsweise durch nummerische, sprachliche und bildhafte Inhalte wie Konzentrationsfähigkeit oder Bearbeitungsgeschwindigkeit. Neben kognitiver Leistungsfähigkeit können aber auch Wissensaspekte wie Mathematikkenntnisse, technisches Verständnis, Rechtschreibung und Grammatik, sprachliche oder fachspezifische Fähigkeiten abgefragt werden. Beide Arten von Tests lassen Aussagen darüber zu, ob Bewerbende besser oder schlechter abschneiden. Demgegenüber stehen berufsbezogene Persönlichkeitstests, welche zur Überprüfung der unternehmenskulturellen Passung (Cultural Fit) dienen (Verhoeven, 2020). In Branchen mit starkem sozialen Bezug, wie beispielsweise in der Medizin (z.B. Lievens, 2013) oder bei Lehrberufen (z.B. Klassen, Kim, Rushby & Bardach, 2019), werden hingegen häufiger nicht-akademische, soziale Kompetenzen wie Resilienz, Empathie und Teamfähigkeit oder auch die interkulturelle Kompetenz (Schnabel, Kelava, Seifert & Kuhlbrodt, 2014) von Bewerbenden getestet. Gerade bei zwischenmenschlichen Kompetenzen hilft die gesteigerte Interpretationsobjektivität und reduzierte Fehleranfälligkeit von Online Assessments bei der Informationsverarbeitung (Fellner, 2019). Bei allen genannten Informationsquellen, Verfahren und Merkmalen im Rahmen einer Fremdselektion geht es in erster Linie darum, die Bewerbenden zu testen und weniger darum, konkrete Informationen zur Stelle oder Unternehmung preiszugeben. Die Ergebnisse im Rahmen einer Fremdselektion werden schlussendlich durch die Personalverantwortlichen einer Unternehmung abgewogen und es wird vom Unternehmen eine Entscheidung bezüglich der Passung getroffen.

2.2.3 Selbstselektion

Aus dem anfänglich beschriebenen Personalbeschaffungsprozess wird deutlich, dass die Auswahlentscheidung ein beidseitiger Prozess ist. So fällt die erste Entscheidung nicht

etwa die Unternehmung nach Eingang einer Bewerbung, sondern vielmehr die Stelleninteressierten, indem sie die Bewerbung an das betreffende Unternehmen abschicken (Diercks, 2021). Ziel jeder Unternehmung sollte sein, die Qualität des Bewerbenden-Pools möglichst hoch zu halten. Es sollen sich im Optimalfall also nur die Personen bewerben, die auch wirklich zur ausgeschriebenen Stelle und Unternehmung passen. Dazu werden im Rahmen der Attraction-Phase immer häufiger sogenannte Online Self-Assessments in Form von Selbsttests oder Matching-Tools eingesetzt. Stelleninteressierte haben dabei die Möglichkeit, noch vor der eigentlichen Bewerbung ihre Passung mit der zu besetzenden Stelle selbständig zu überprüfen und zu entscheiden, ob eine Bewerbung überhaupt sinnvoll ist. Der Unterschied zu klassischen Online Assessments liegt darin, dass das Testergebnis nur den Stelleninteressierten angezeigt wird. Zudem haben Online Self-Assessments eine gewisse «Anlockfunktion» (Diercks, 2021). Sie sollen das Interesse von potenziellen Bewerbenden wecken, neugierig machen und mit realistischen Informationen über den Beruf oder die Unternehmung dabei helfen, Vorurteile und überzogene Erwartungen zu relativieren und Chancen aufzuzeigen. So werden Online Self-Assessments in der Literatur sogar manchmal im Bereich vom Personalmarketing verortet (Hiltmann, 2013), wobei mithilfe von spielerischen Elementen (siehe Gamification weiter unten) sowohl die eigenen Stärken und Schwächen, als auch das Unternehmen besser kennengelernt werden können. Laut Ott et al. (2017) werden Informationen zur ausgeschriebenen Stelle und zum Unternehmen durch die Interaktivität mit solchen Tools intensiver verarbeitet als durch eine reine Auflistung auf Karrierewebseiten. Zwar ist diese Art von Interaktivität in der Personalauswahl noch verhältnismässig neu, dennoch gibt es bereits einige Beispiele, wie Unternehmen solche Selbstselektion-Tools zur Verfügung stellen. So sind beispielsweise bei der Firma Swisscom auf der Webseite bei vielen Stelleninseraten sogenannte «Skill Checks» aufgeschaltet. Dort können Stelleninteressierte durch Anwählen von verschiedenen Angaben zu ihrer eigenen Ausbildung und ihren Fähigkeiten testen, ob ihre Angaben zur

ausgeschriebenen Stelle passen oder nicht. Falls die Angaben passen, erscheint direkt die Aufforderung zur Bewerbung (Swisscom, 2022). Diese Art von Selbstselektions-Tool wird *Matching-Tool* genannt. Dabei geht es weniger um eine exakte Messung, sondern vielmehr um eine Entscheidungshilfe. Etwas umfangreichere Selbstselektions-Tools sind vor allem in der Berufs- und Studienwahl schon sehr verbreitet. Ein etabliertes Beispiel ist der Online Selbsttest Psychologie der Universität Zürich und Fachhochschule Nordwestschweiz (2022), der Studieninteressierten dabei hilft, ihre Interessen und Fähigkeiten besser kennenzulernen, sich mit den Anforderungen des Studiengangs Psychologie vertraut zu machen und eine Entscheidung für oder gegen das Studium zu treffen.

Gemäss Katzlinger (2017) führen Online Self-Assessments zu einem angenehmeren Klima bei Jobinterviews, weil die Bewerbenden schon etwas über die Firma wissen. Ausserdem reduzieren sie die Unsicherheit der Bewerbenden und erhöhen die empfundene Fairness des Auswahlverfahrens. Jedoch ist bezüglich letzterem gerade bei der Selbstselektion Vorsicht geboten. Die empfundene Fairness und damit auch die eingeschätzte Attraktivität eines Unternehmens hängen auch von der Rückmeldung eines Selbstselektions-Tools ab. Anseel und Lievens (2009) konnten in ihrer Studie zeigen, dass der Zusammenhang zwischen der Rückmeldung (niedrige vs. hohe Passung) und der empfundenen Attraktivität einer Unternehmung von der Akzeptanz der erhaltenen Rückmeldung mediiert wird. Also je positiver die Rückmeldung, desto positiver die Einschätzung der Attraktivität der Unternehmung. Um dabei nicht den eigentlichen Zweck des Tools, nämlich die Selbstselektion, zu untergraben, soll die Rückmeldung von Selbstselektionsinstrumenten möglichst konkret und anforderungsbezogen formuliert werden, damit die Plausibilität der selbstbezogenen Informationen ausreichend verarbeitet werden und dadurch dir richtige Entscheidung getroffen werden kann (Anseel & Lievens, 2009).

2.2.4 Darbietungsweise und Faking

Ein anderer wichtiger Aspekt von Online (Self-)Assessments ist die Darbietungsweise. Der Begriff *online* wird häufig mit *ohne Aufsicht* gleichgesetzt. Dies ist aber nicht ganz richtig. Online heisst lediglich, dass die Testung an einem Computer oder mobilen Gerät stattfindet, das mit anderen Geräten via Internet oder Intranet verbunden ist. Dabei ist eine Überwachung der Testung nicht ausgeschlossen. Ist ein Test hingegen online für die Öffentlichkeit frei zugänglich, so spricht man von einer unbeaufsichtigten Testung (The International Test Comission, 2006). Grundsätzlich werden aber Online (Self-)Assessments im Vergleich mit klassischeren Auswahlverfahren häufig in unbeaufsichtigter Umgebung (von zu Hause aus) durchgeführt, was bezogen auf die zeitliche und ortsunabhängige Flexibilität häufig als Vorteil genannt wird. Gleichzeitig entstehen dadurch unterschiedliche Herausforderungen. Zum einen kann nicht kontrolliert werden, unter welchen ortsabhängigen und motivationalen Bedingungen (z.B. Ablenkung) der Test durchgeführt wurde, was zu Verzerrungseffekten führen kann. Ausserdem kann es zum Einsatz unerlaubter Hilfsmittel oder anderweitigem Faking (engl. für Vortäuschen oder Fälschen) kommen. Beispielsweise könnten andere (geeignetere) Personen den Test für die bewerbende Person durchführen oder es kann zu einem unkontrollierten Ausmass an Übung kommen, wenn ein Test mit unterschiedlichen Identitäten mehrmals durchgeführt wird (Ott et al., 2017). Karim, Kaminsky und Behrend (2014) konnten in ihrer Studie zeigen, dass die Online-Aufsicht via Webcam während einer Testung zu weniger Faking führte, sich aber auch negativ auf die Reaktion der Teilnehmenden auswirkte. Die Aufsicht hatte jedoch keinen direkten Einfluss auf die Testleistung. Während andere Autorinnen und Autoren beispielsweise auf anschliessende Überprüfungstests setzen (Lievens & Burke, 2011), verweisen Landers und Sackett (2012) auf den grossen Vorteil eines grösseren Bewerberpools bei Online-Formaten, der den Nachteilen von möglichem Faking überwiegt.

2.2.5 Datenschutz & Qualität

Wie bereits in den vorangegangenen Kapiteln geschildet, kann im hartumkämpften Arbeitsmarkt eine möglichst effiziente, valide und schnelle Personalauswahl entscheidend sein. Dabei rücken oftmals wichtige Aspekte in den Hintergrund, die aber ebenfalls einen nicht unbedeutenden Einfluss auf die Bewerbenden haben kann. Der wahrgenommene Datenschutz ist beispielsweise umso wichtiger, je grösser der Testanteil bei einem Online Assessment ist (Ott et al., 2017). Ausserdem bewerben sich Stelleninteressierte eher, wenn das dafür vorgesehene Bewerbungsformular als sicher wahrgenommen wird (Bauer et al., 2006). Fellner (2019) weist in diesem Zusammenhang darauf hin, dass allgemein immer mehr Daten über Bewerbende zum Analysieren und Auswerten für Unternehmen zur Verfügung stehen. Zwar war die Eignungsdiagnostik schon immer sehr datenintensiv, doch mit der digitalen Erfassung von beispielsweise Maus- oder Blickbewegungen sowie Stimm- und Sprachanalysen während der Nutzung eines digitalen Formats kommen neue Dimensionen dazu. Umso wichtiger ist es, dass sichergestellt wird, dass die Daten von guter Qualität sind und nur im Rahmen von Datenschutzbestimmungen zum gemeinsamen Nutzen von Bewerbenden und der Unternehmung verwendet werden (Fellner, 2019). Solche Qualitätsstandards für computergestützte Tests werden beispielsweise in den International Guidelines on Computer-Based and Internet-Delivered Testing von The International Test Comission (2006) geregelt. Dabei werden auch die technischen Voraussetzungen und die Frage danach, wer für die «automatischen» Diagnosen von Online Assessments verantwortlich ist, geregelt. Ausserdem sollten bei der Entwicklung und Bewertung von Online Assessments auch die Qualitätsanforderungen der DIN 33430 (2016) berücksichtigt werden, welche die Anforderungen an berufsbezogene Eignungsdiagnostik formulieren.

2.2.6 Gamification

Gamification wird als Konzept verstanden, das verschiedene Techniken aus dem Game-Design nutzt, um die Effektivität von bestehenden Methoden zu verbessern und

ihnen gleichzeitig einen spielerischeren Charakter zu verleihen (Armstrong, Ferrell, Collmus & Landers, 2016). Dabei werden beispielsweise Elemente wie Zielfokussierung, Belohnungsmechanismen und Fortschrittsüberwachung als Mechanismen genutzt (Glover, 2013), um das Engagement, die Motivation und Performance zu steigern (Larson, 2019). Doch auch wenn bei der Gamification Game-Elemente genutzt werden, besteht ein entscheidender Unterschied zu klassischen Spielen. Bei der Gamification einer Methode liegt immer ein unternehmerisches Ziel zu Grunde, wie beispielsweise die Steigerung der Effektivität von den Nutzenden (Singh, 2012). Larson (2019) untersuchte in seinem Review die Anwendung von Gamification im Arbeitskontext. Dabei sieht er bei der Anwendung solcher game-basierter Online Assessments in der Arbeitswelt vor allem Chancen in den Bereichen Innovation, Ausbildung und Schulung, Leistungssteigerung und Produktivität sowie Mitarbeitergewinnung und -bindung. Der Anwendungsschwerpunkt liegt gemäss Larson (2019) im Bereich Motivation, Rekrutierung und Schulung, wobei vor allem im Bereich Schulung bereits signifikante, wissenschaftliche Evidenz für den Einsatz solcher game-basierter Assessments gefunden werden konnte. Beispielsweise wurden in einer Studie 140'000 Mitarbeitende der amerikanischen Firma Walmart geschult, unter anderem auch im Bereich der sozialen Kompetenzen. Dabei berichteten die mit VR geschulten Verkaufsmitarbeitenden nicht nur 30% mehr Zufriedenheit mit der Schulung, 70% von ihnen zeigten schlussendlich auch bessere Leistungen als diejenigen, die mit klassischen Schulungsmethoden geschult wurden (Rogers, 2019).

In der Rekrutierung kommt Gamification zwar ebenfalls schon zum Einsatz, jedoch gibt es in diesem Bereich noch wenig Forschung zur Effektivität (Georgiou, Gouras & Nikolaou, 2019). In seinem Review zur Anwendung von Gamification im Arbeitskontext konnte Kalafatoglu (2020) nur zwei Studien zum Einsatz von Gamification-Elementen im Personalbeschaffungsprozess. Diercks (2021) nennt in diesem Zusammenhang den Begriff «Recrutainment. Ein möglicher Vorteil von game-basierten Assessments bei der Rekrutierung könnte sein, dass die Aufmerksamkeit der Individuen durch den spielerischen

Charakter von der eigentlichen Bewerbungssituation abgelenkt wird und dadurch das tatsächliche Verhalten zum Vorschein kommt. Dies würde den Einfluss von sozial erwünschtem Verhalten verringern und dadurch die Vorhersage der Job Performance verbessern (Georgiou et al., 2019). Ausserdem erwarten Fetzer, McNamara und Geimer (2017) durch den Einsatz von game-basierten Selektionsmethoden grösseres Engagement und ein positiveres Image der betreffenden Arbeitgebenden. Letzteres konnten Gkorezis, Georgiou, Nikolaou und Kyriazati (2021) in ihrer Studie bestätigen. Die Autoren konnten zeigen, dass ein game-basierter SJT im Vergleich zu einem traditionellen SJT eine signifikant positiver wahrgenommene Attraktivität des Unternehmens bewirkte und dadurch indirekt auch einen positiven Einfluss auf die Empfehlungsabsichten der Befragten hatte. Katzlinger (2017) analysierte die Gemeinsamkeiten und Unterschiede aktueller Online Self-Assessments mit Game-Elementen. Alle berücksichtigten Tools wurden von Firmen genutzt mit einem Umsatz von mehr als einer Milliarde Euros Umsatz, mehr als 10'000 Beschäftigten und die Simulationen waren für Absolventen, Studierende und Praktikanten ausgerichtet. Die Autorin analysierte insgesamt 15 Instrumente und konnte zeigen, dass folgende Schlüssel-Features von game-basierten Online Self-Assessments einen signifikant positiven Einfluss auf das Engagement der Nutzenden hatten: Schnellere Feedback-Schleifen, klare Ziele und Regeln, spannende Hintergrundgeschichte und herausfordernde, lösbare Aufgaben. Ein Beispiel der analysierten Studien war ein französisches Postunternehmen, das die Stelleninteressierten eine Routine-Woche einer Postkarriere erleben liess. Ziel war es, die Erwartungen bei der Nutzung steuern zu können und nur wirklich motivierte Bewerbungen zu bekommen. Das Resultat war eine Reduktion der Abbruchrate im anschliessenden Bewerbungsprozess von 25% auf 8% und die Bewerbenden kamen zudem besser informiert ins Gespräch (Ireland, 2016). In einer anderen Studie zur Überprüfung der Validität von game-basierten Assessments entwickelten Georgiou et al. (2019) zuerst einen textbasierten SJT zur Messung der sozialen Kompetenzen Resilienz, Anpassungsfähigkeit, Flexibilität und

Entscheidungsfreudigkeit mit einer bestätigten Konstruktvalidität (konvergente und diskriminante Validität). In einem zweiten Schritt hat die Autorenschaft durch Hinzufügen von Game-Elementen wie einer Abenteuer-Story, Avataren, Feedback sowie visuellen und stimmlichen Erzählungen einen game-basierten Version des SJT entwickelt. Zwar waren die Voraussetzungen für die statistischen Berechnungen nicht überall ideal, es konnte jedoch die Konstruktvalidität des game-basierten SJT mit einigen Vorbehalten bestätigt werden. Damit zeigten die Autoren laut eigener Aussage zum ersten Mal überhaupt die psychometrischen Eigenschaften eines game-basierten Selektions-Instruments und befürworten damit die Möglichkeit, dass game-basierte Assessments zur Vorhersage von späterem Arbeitsverhalten beitragen können (Georgiou et al., 2019).

2.3 Der Situational Judgment Test

Wie bereits im vorherigen Kapitel angedeutet, können Situational Judgment Tests in Form eines Online Assessments zur Einschätzung der Eignung von Bewerbenden verwendet werden. Der SJT ist ein etabliertes Verfahren zur prädiktiven Einschätzung des späteren Leistungsverhaltens einer Person und ist deshalb auch ein beliebtes Instrument im Rahmen des Personalbeschaffungsprozess. In diesem Kapitel soll zuerst auf die Besonderheiten des SJT als simulationsorientiertes Verfahren eingegangen werden, bevor die Konstruktion eines SJT genauer beschrieben wird. Nebst unterschiedlicher Gestaltungsformen soll ein besonderes Augenmerk auf die prädiktive Validität dieses Verfahrens gelegt werden.

2.3.1 Simulationsorientiertes Verfahren

Situational Judgment Tests werden seit Anfang der 90er-Jahre zu Zwecken der Personalauswahl vermehrt erforscht. Der SJT gilt als simulationsorientiertes Verfahren und wird häufig eingesetzt, da er ein Stück aus dem spezifischen Berufsalltag in die durch diagnostische Verfahren und abstrakte Kompetenzmerkmale geprägte Umgebung des Selektionsprozess bringt (Muck, 2013). Der Prototyp von simulationsorientierten Verfahren

ist die klassische Arbeitsprobe, die es vermutlich schon so lange gibt, wie die Personalauswahl insgesamt. Dabei werden Bewerbende in die Unternehmung eingeladen, um konkrete Aufgaben aus dem zukünftigen Berufsalltag auszuführen, sodass sich die interessierenden Fertigkeiten direkt beobachten und bewerten lassen. Über die Jahre und mit der Digitalisierung sind zahlreiche Abwandlungen der klassischen Arbeitsprobe entwickelt worden, wie beispielsweise das Assessment Center, computergestützte Szenarios, situative Interview- und Fragebogentechniken, sowie der Situational Judgment Test (Kanning & Schuler, 2014). Weekley et al. (2015) definieren eine Arbeitssimulation als Rekonstruktion der für die Arbeit nötigen Anforderungen, um die Kapazität von Bewerbenden zur Erfüllung dieser Anforderungen zu messen. Die gezeigten Verhaltensweisen der Bewerbenden werden dabei verglichen mit einem Ideal (Weekley et al., 2015). Beim SJT werden den Bewerbenden möglichst explizite und realitätsnahe Situationen aus dem Berufsalltag (Stimuli) in Form von Texten oder Videoclips präsentiert. Die Bewerbenden sollen dann entscheiden, wie sie in der beschriebenen Situation reagieren würden. Das Antwortformat (Response) kann als eine Auswahl an vorgegebenen Antwortalternativen bzw. Verhaltensweisen präsentiert werden oder auch offen gelassen werden (Kanning & Schuler, 2014).

Ein typisches Merkmal von simulationsorientierten Verfahren ist, dass sie oftmals nicht ein spezifisches Konstrukt messen, sondern eher ein beispielhaftes Verhalten der Bewerbenden in einer Situation zeigen (Kanning, 2009; Muck, 2013). Deshalb sind Simulationen als sehr jobspezifisch zu bewerten und oftmals nicht für verschiedene Stellen oder Unternehmen kopierbar (Weekley et al., 2015). Um simulationsorientierte Verfahren zu systematisieren, wird ihr Realitätsbezug bewertet. In ihrem Review zu sogenannten «low-fidelity» Simulationen (engl. Simulationen mit geringem Realitätsbezug) nehmen Weekley, et al. (2015) die in Abbildung 3 dargestellte Einordnung vor. Eine eher realitätsnahe Stimulus-Komponente könnte beispielsweise eine Simulation innerhalb eines Rollenspiels im Assessment Center oder ein Video sein, wohingegen eine textbasierte Beschreibung

einer Situation eine geringe Realitätsnähe aufweist. Beim Antwortformat ist eine geringe Realitätsnähe gegeben, wenn die Bewerbenden aus einem Set von vorgegebenen Antwortalternativen die passendste auswählen können. Wenn Bewerbende jedoch direkt auf einen präsentierten Stimulus reagieren sollen, ohne vorgegebene Alternativen, so wird die Realitätsnähe höher bewertet (Weekley et al., 2015).

		Response-Komponente	
		Low fidelity	High fidelity
Stimulus-Komponente	Low fidelity	Text SJT	Situational Interview
	High fidelity	Multimedia SJT	Assessment Center

Abbildung 3. Stimulus- und Response-Komponente nach Weekley et al. (2015)

Das Assessment Center schneidet bei dieser Bewertung am besten ab, da sowohl Stimulus- als auch Response-Komponente als realitätsnah eingeschätzt werden. Ein Beispiel wäre ein Rollenspiel bei dem eine Arbeitssituation simuliert wird, auf die die bewerbende Person sofort selbständig reagieren muss. Textbasierte SJT hingegen schneiden bezüglich Realitätsbezug am schlechtesten ab, da der schriftliche Stimulus Raum für das Vorstellungsvermögen der Befragten lässt und das Antwortformat vorgegeben ist. Multimedia-basierte SJT heben sich mit einem höheren Realitätsbezug der Stimulus-Komponente von dem textbasierten SJT ab (Weekley et al., 2015). Obwohl sowohl text- also auch multimediabasierte SJT wegen den vorgegebenen Antwortalternativen gegenüber anderen Verfahren schlechter abschneiden, werden sie in der Praxis sehr häufig genutzt. Dies liegt laut Weekley et al. (2015) daran, dass vorgegebene Antwortalternativen viel weniger aufwendig im Auswertungsaufwand und deshalb ökonomischer in der Auswertung sind im Vergleich mit offenen Antwortmöglichkeiten und dadurch auch grössere Mengen von Bewerbenden eingeschätzt werden können. Ausserdem können auch Situationen simuliert werden, die sich kaum real nachstellen lassen, da sie zu gefährlich sind (Kanning & Schuler, 2014), wie beispielsweise mit angehenden Polizeibeamten (Leeds et al., 2003).

Eine andere, eher selten gebrauchte Möglichkeit, ein Verfahren noch realistischer zu gestalten, sind interaktive SJTs. Dabei stehen Items (also einzelne Stimuli oder Fragen) nicht isoliert nebeneinander, sondern sind miteinander verknüpft. Das heisst, je nachdem welche Antwortalternative durch die Bewerbenden ausgewählt wird, wird ein anderes nachfolgendes Item präsentiert, damit der geschichtliche Verlauf Sinn macht. Die Geschichte geht also nach der Wahl einer unpassenden Antwortalternative anders weiter, als wenn Bewerbende wie gewünscht reagieren. Zwar nimmt dabei der Realitätsbezug deutlich zu, jedoch entsteht ein sehr viel höherer Entwicklungsaufwand, weshalb diese Variante eher selten gewählt wird. Während bei zwei nicht verknüpften Items mit jeweils vier Antwortalternativen insgesamt zehn Szenarien (zwei Stimuli und acht Antwortalternativen) entwickelt werden müssten, sind es bei zwei verknüpften Items bis zu 21 zu entwickelnde Szenarien (Kanning & Schuler, 2014).

2.3.2 Virtual Reality und 360°-Videos

Mit der heutigen Technologie ist es möglich, bezüglich der Realitätsnähe von Simulationen noch einen Schritt weiterzugehen. Mithilfe von Virtual Reality (VR) können computergenerierte, dreidimensionale Umgebungen geschaffen werden, mit denen als 3D-Video auf einem Computer (nicht-immersiv) oder in sogenannten VR-Brillen (immersiv) interagiert werden kann (Preuss & Kauffeld, 2019). Während betrachtende Personen bei der immersiven VR durch die Interaktion mit VR-Brille, Sensorhandschuhe und weitere Sensoren zur Messung von körperlichen Aktivitäten komplett in die virtuelle Welt eintauchen können, ist das Erlebnis bei 3D-Videos am Computer durch die Interaktion mit Tastatur und Maus etwas weniger realitätsnah. Eine weitere Variante ist die Augmented Reality (AR), bei der computeranimierte Elemente durch das VR-System in die echte Welt übertragen werden (Lee & Wong, 2008).

Es gibt bereits einige Forschung zur Wirksamkeit von immersiven VR-Simulationen zur Unterstützung von Schulungen. Dies würde im Personalbeschaffungsprozess vor allem das On-Boarding betreffen. Hamilton et al. (2020) konnten dazu in ihrem Review zeigen,

dass von den insgesamt 29 berücksichtigten Studien rund die Hälfte signifikant positive Lerneffekte mit immersiven VR angeben. Als Alternativmethoden wurden computerbasierte Simulationen oder Powerpoint-Präsentationen verglichen. Webster (2016) erklärt sich die positiven Effekte von immersiven VR beim Lernen vor allem durch das vollständige Abtauchen in die virtuelle Welt (Realitätsnähe) und das gleichzeitige Ansprechen von verschiedenen Sinnessystemen (visuell, auditiv, haptisch). Adarve-Gómez, Castillo-Carvajal, Restrepo-Zapata und Villar-Vega (2019) sehen den grössten Vorteil hingegen darin, dass die Mitarbeitenden in einer sicheren Umgebung verschiedene Verhaltensalternativen ausprobieren können und direkt aus der Analyse der Konsequenzen lernen. Ausserdem weisen die Autoren auf das grosse zeitliche und finanzielle Einsparungspotenzial beim Einsatz von VR hin.

Betrachtetet man die früheren Phasen des Personalbeschaffungsprozess gibt es ebenfalls bereits Entwicklungen, die sich VR-Technologie zur Hilfe nehmen. Fominykh und Prasolova-Førland (2019) entwickelten ein neues Konzept, welches jungen Arbeitssuchenden ermöglichen sollte, mithilfe von immersiven VR verschiedene Berufe «auszuprobieren». Die Autoren trugen dabei dem Umstand Rechnung, dass junge Personen beim Eintritt in die Berufswelt häufig noch nicht wissen, was sie arbeiten wollen und viele mögliche Berufsprofile auch gar nicht kennen. Gleichzeitig sollte den Arbeitgebenden ein Tool zur Seite gestellt werden, in dem sie die Karrieremöglichkeiten in ihrer Unternehmung präsentieren konnten. Bei ihrer Befragung nach den Bedürfnissen von jungen Arbeitssuchenden stellten die Autoren fest, dass Praktika oftmals nicht den erwarteten Nutzen gebracht haben, da typische Berufssituationen gar nicht darin vorkommen. Auch textüberladene Jobbeschreibungen werden als wenig hilfreich angesehen. Die befragten Personen wünschten sich neben Bewerbungsgespräch-Simulationen vor allem die tatsächliche Veranschaulichung von Arbeitsplätzen, Arbeitsaufgaben sowie den dazu erforderlichen Kompetenzen und Feedback zu ihrer Leistung. Auf eine Wettbewerbssituation oder ein ortsgebundenes Setting können sie

hingegen verzichten. Mit diesem Wissen entwickelten Fominykh und Prasolova-Førland (2019) sowohl die FisheryVR-App, bei der der Alltag in einem Fischzuchtbetrieb erlebt werden kann, als auch die InterviewVR-App, bei der ein immersives Bewerbungsgespräch durchlebt werden kann. Beide Apps wurden von unterschiedlichen Anspruchsgruppen getestet und Feedbacks wurden gesammelt. Die durchwegs positiven Rückmeldungen zeigen laut der Autorenschaft, dass das Konzept der immersiven Arbeitsprobe den jungen Arbeitssuchenden nicht nur einen besseren Einblick in den Berufsalltag geben kann, sondern auch eine motivierende, kosteneffektive, leichter zugängliche und besser akzeptierte Alternative zu klassischen Praktika und anderen Berufsberatungsangeboten ist (Fominykh & Prasolova-Førland, 2019).

2.3.3 Entwicklung eines Situational Judgment Test

Mit den unterschiedlichen Gestaltungsmöglichkeiten der Stimulus- und Response-Komponenten gleicht kein SJT dem anderen. Zur Entwicklung von SJTs sind in jedem Fall einige Schritte zu durchlaufen und einige Entscheidungen zu treffen. Muck (2013) fasst den Prozess der SJT-Entwicklung gemäss Abbildung 4 in vier aufeinander folgende Schritte zusammen.

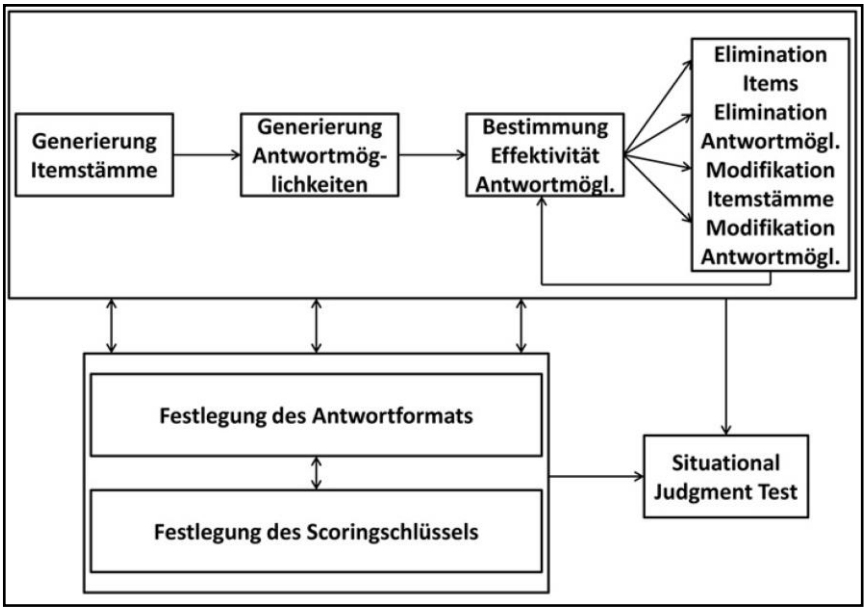

Abbildung 4. Entwicklung eines SJT nach Muck (2013)

Nachdem zuerst Itemstämme und Antwortmöglichkeiten e generiert werden, müssen im Anschluss die Effektivität der Antwortmöglichkeiten bestimmt und wo nötig Items modifiziert oder ausgeschlossen werden. Ausserdem wird im Laufe des Entwicklungsprozesses entschieden werden, wie das Antwortformat formuliert und der Scoringschlüssel definiert werden sollen. In den folgenden Abschnitten werden die einzelnen Schritte sowie Überlegungen dazu beschrieben.

Generierung der Itemstämme. Der Itemstamm besteht bei einem SJT aus einer konkreten Situationsbeschreibung. Grundsätzlich unterscheidet Muck (2013) zwei Herangehensweisen, um Itemstämme zu identifizieren. Zum einen können mit der Critical Incident Technique (CIT) von Flanagan (1954) erfolgskritische (typische und wichtige) Arbeitssituationen in einem spezifischen Kontext gefunden werden. Dazu sollen Fachleute aus dem entsprechenden Arbeitsbereich (sogenannte subject matter experts; SME) wie beispielsweise Vorgesetzte, Mitarbeitende mit gleichen Aufgaben oder auch die Kundschaft an vergangene kritische Situationen denken, die durch bestimmte Verhaltensweisen zu einer besonders effektiven oder ineffektiven Aufgabenerfüllung geführt haben (Blickle, 2019). Soll der SJT hingegen eine Theorie oder ein Konstrukt abbilden, so können die Situationen für den SJT auch durch den Testentwickler oder die Testentwicklerin selbst generiert werden. So wird sichergestellt, dass alle Aspekte der betreffenden Theorie abgedeckt werden (Weekley, Ployhart & Holtz, 2006). Gemäss Muck (2013) ist auch die Kombination beider Herangehensweisen möglich. Bei der Formulierung der Situationen ist unter anderem darauf zu achten, dass die Beschreibung nicht zu lang oder zu schwer zu verstehen ist. Sind die Items zu komplex formuliert, so korrelieren die Ergebnisse eher mit der allgemeinen Intelligenz der Befragten als mit dem abgefragten Konstrukt (Muck, 2013). Ausserdem sollen die Situationen so formuliert sein, dass ein gewisses Dilemma für die Befragten entsteht (beispielsweise mehr Zeitaufwand oder eine unangenehme Konfrontation). Dadurch kann zwischen Personen mit niedrigen und hohen Ausprägungen bei der für die Situationsbewältigung bedeutsamen Kompetenzen unterschieden werden

(Bledow & Frese, 2009). Bei Bedarf werden die generierten Itemstämme zum Schluss ausgewählten Kompetenzbereichen zugeordnet (Garman, Johnson & Howard, 2006). Muck (2013) betont dabei die Wichtigkeit, dass es sich um andere Fachleute für die Zuordnung handeln soll, als diejenigen, die die Items formuliert haben. Dieser Schritt entfällt jedoch, wenn alle Items das gleiche Konstrukt oder einen gemeinsam abzubildenden Inhaltsbereich repräsentieren.

Generierung der Antwortmöglichkeiten. Gemäss Weekley et al. (2006) werden normalerweise mindestens drei bis zwölf Antwortmöglichkeiten pro Itemstamm generiert. Die Realitätsnähe nimmt mit steigender Anzahl Items zu und damit auch der Konstruktionsaufwand. Zur Generierung der Antwortmöglichkeiten stehen ähnliche Vorgehensweisen wie bei Generierung der Itemstämme zur Auswahl: Zum einen können SME aufgrund ihrer Erfahrung mögliche Antwort- bzw. Verhaltensalternativen bezüglich Realitätsnähe und Sinnhaftigkeit am besten abschätzen (McDaniel & Nguyen, 2001). Andererseits könnten auch unerfahrene Personen befragt werden, um eine grössere Bandbreite an Antwortmöglichkeiten zu gewinnen (Muck, 2013). Sollen hingegen mit dem SJT bestimmte Konstrukte erfasst werden, so muss auch bei der Generierung der Antwortalternativen darauf geachtet werden, dass sie das Konstrukt vollumfänglich repräsentieren (Ployhart & MacKenzie, 2011).

Festlegung des Antwortformats. Unter dem Antwortformat wird die Formulierung der Instruktion verstanden, die Einfluss auf die Gestaltung der Antwortalternativen und die spätere inhaltliche Interpretation hat. Es muss festgelegt werden, ob die Befragten angeben sollen, wie man sich gemäss ihrer Meinung in der entsprechenden Situation verhalten sollte (Should-do-Instruktion) oder wie die Befragten selbst sich in der entsprechenden Situation verhalten würden (Would-do-Instruktion). Should-do-Instruktionen korrelieren dabei stärker mit der Intelligenz der Befragten (Nguyen, Bidermann und McDaniel, 2005) und zeigen höhere Korrelationen mit deren kognitiven Fähigkeiten (McDaniel, Harman, Whetzel und Grubb, 2007). Ausserdem wird davon ausgegangen, dass Should-do-Instruktionen weniger

anfällig sind für Verfälschung (Nguyen et al., 2005), da es bei wissensbasierten Fragestellungen eher darum geht, die *richtige* Antwort zu identifizieren, also die *maximale* Leistung zu erreichen. Testergebnisse mit der Would-do-Instruktion zeigen hingegen eher die *typische* Leistung (McDaniel et al., 2007), hängen stärker mit Persönlichkeitsskalen zusammen und haben deshalb eher einen verhaltensbezogenen Charakter (Nguyen et al., 2005). Ebenfalls abhängig vom gewählten Antwortformat ist die Schwierigkeit einer Befragung. Diese lässt sich steigern, wenn die Befragten nicht nur die beste oder für sie passendste Verhaltensalternative auswählen sollen (Single-Choice), sondern die beste *und* die schlechteste Verhaltensalternative (Multiple Choice). Ein weiteres mögliches Antwortformat wäre, alle Alternativen in eine Rangreihe bringen zu lassen (= Ranking; McDaniel & Nguyen, 2001) oder für jede Alternative auf einer mehrstufigen Likert-Skala einschätzen zu lassen, wie geeignet sie ist oder wie wahrscheinlich es ist, dass die Befragten dieses Verhalten zeigen würden (= Rating). Ein Rating jeder Antwortalternative hat gegenüber Forced Choice-Formaten den Vorteil, dass die Ergebnisse eine höhere Varianz aufweisen. Dadurch wird auch die Reliabilität und Validität des betreffenden SJT positiv beeinflusst (Ployhart & Ehrhart, 2003).

Bestimmung der Effektivität der Antwortmöglichkeiten und Scoring. Hierbei geht es in erster Linie darum, die gefundenen Antwortalternativen bezüglich ihrer Effektivität in der entsprechenden Situation zu bewerten (Bergman, Drasgow, Donovan, Henning & Juraska, 2006). Grundsätzlich sind in der Literatur drei verschiedene Scoring-Methoden zu finden. Als eine vierte Alternative nennt Muck (2013) die hybride Variante, eine Kombination der drei Methoden. Welches Scoring-Verfahren schlussendlich gewählt werden soll, ist abhängig von theoretischen und praktischen Überlegungen bezüglich der geplanten Anwendung (Bergman et al., 2006).

Eine der gängigsten Scoring-Methoden ist das *expertenbasierte* Scoring. Dabei sollen wiederum SME mit substanziellem Wissen die Antwortalternativen gemäss ihrer Relevanz für ein bestimmtes Kriterium wie beispielsweise die Arbeitsleistung bewerten. Da die

meisten SJT komplexe, soziale oder praktische Verhaltensweisen simulieren, die aus verschiedenen Perspektiven angeschaut werden können, macht es Sinn, die Effektivität in der entsprechenden Situation beurteilen zu lassen. Jedoch besteht bei diesem Vorgehen das Risiko, dass bei der Bewertung auf Firmennormen oder firmenspezifische Vorgaben zurückgegriffen wird, welche Bewerbende gar nicht wissen können (McDaniel, Whetzel, & Nguyen, 2006).

Beim *empirischen* Scoring wird die Einstufung der Antwortalternativen eines Items anhand ihres Zusammenhangs mit einem Kriterium wie beispielsweise der Arbeitsleistung gewichtet. So sollen die Antwortalternativen gut zwischen leistungsstarken und leistungsschwachen Mitarbeitenden unterscheiden. Dabei werden die zu erfüllenden Bedingungen, wie beispielsweise die Korrelationshöhe einer Antwortalternative mit dem Kriterium, im Vorfeld festgelegt (Bergman et al., 2006). Weekely und Jones (1997) haben beispielsweise diese Methode für ihren videobasierten SJT zur Selektion von Verkaufsmitarbeitenden genutzt. Dazu haben 684 Mitarbeitende den SJT ausgefüllt und wurden ausserdem von ihren Vorgesetzten in ihrer Arbeitsleistung bewertet. Die Antwortalternative mit der höchsten durchschnittlichen Arbeitsleistung wurde dann als die effektivste Antwortalternative im SJT bewertet.

Beim *theoretischen oder konstruktbasierten* Scoring orientiert sich die entwickelnde Person zur Bestimmung der effektivsten oder ineffektivste Antwortalternative an dem theoretischen Konzept (Bergman et al., 2006). Das Vorgehen bei der Konstruktion ist dabei vergleichbar mit der Entwicklung einer Persönlichkeitsskala durch eine Skalenanalyse (Mumford, 1999). Aufgrund ihrer höheren Transparenz sind theoretische Scorings laut Hough und Paullin (1994) jedoch anfälliger für Faking.

Festlegung des Scoring-Schlüssels. Bergman et al. (2006) beschreiben die Bestimmung eines passenden Auswertungsschlüssels für einen SJT im Titel ihres Artikels sehr treffend: «Scoring situational judgment tests: Once you get the data, your troubles begin». Denn im Gegensatz zu vielen anderen Instrumenten für die Personalauswahl gibt

es bei SJT meist keine objektiv richtige Antwort. Bei der Testentwicklung muss deshalb auch entschieden werden, auf welche Art und Weise die Antworten der befragten Personen mit der Experteneinschätzung oder Theorie verglichen wird (McDaniel, Psotka, Legree, Yost & Weekley, 2011) und wie diese Angaben quantifiziert werden. Bei Single und Multiple Choice-Antwortformaten kann dies zum einen über den Zustimmungsgrad der SME bewertet werden. Beispielsweise haben MacCann und Roberts (2008) den Prozentsatz der SME, die der Antwortalternative zugestimmt haben, als Zahlenwert der Antwortalternative verwendet. Andere Autoren vergeben Punkte für die beste oder schlechteste Antwort oder für beide. Lievens und Peeters (2008) vergeben beispielsweise für die Wahl der besten Antwort +1 Punkte und für die schlechteste Antwort -1 Punkte. Für alle anderen Antworten gibt es 0 Punkte. Bei der Verwendung einer Likert-Skala pro Antwortalternative wird der Score eines Items mittels unterschiedlicher Verrechnung der einzelnen Likert-Antworten ermittelt. Manche Autoren verrechnen die absoluten Werte der besten und schlechtesten Antwortmöglichkeiten (Mumford, Van Iddekinge, Morgeson & Campion, 2008), andere berücksichtigen auch hier die zusätzlich die Effektivität aufgrund des expertenbasierten Ratings (Ployhart & Ehrhart, 2003).

Elimination oder Modifikation von Items und Antwortmöglichkeiten. Muck (2013) fasst verschiedene in der Literatur angewandte Kriterien zusammen, die Hinweise darauf liefern könnten, dass ein Item oder bestimmte Antwortmöglichkeiten nicht funktionieren und deshalb geändert oder vom SJT ausgeschlossen werden müssen. Unter anderem wird die mangelnde Übereinstimmungen der Fachleute bezüglich der Effektivität der Antwortalternativen genannt (Lievens & Patterson, 2011). Andere Kriterien könnten sein, dass eine Antwortalternative zu nahe an der effektivsten Antwort liegt oder zu wenig Varianz im Antwortverhalten generiert werden kann. Falls in der Diskussion der Fachleute keine passende Modifikation gefunden wird, soll das betreffende Item ausgeschlossen werden (Whetzel & McDaniel, 2009).

2.3.4 Prädiktive Validität von Situational Judgment Tests

Eines der wichtigsten Ziele in der Personalauswahl ist es, auf Basis von ausgewählten Informationen zu Bewerbenden ihr späteres Leistungsverhalten im Job (Job Perfomance) vorherzusagen. Eine einheitliche Definition der Job Performance ist in der Literatur aber nicht zu finden. Motowidlo (2003) definiert die Job Performance als den erwarteten Wert, den Mitarbeitende durch ihre Verhaltensweisen über eine bestimmte Zeitperiode für eine Unternehmung erbringen. Job Performance ist demnach ein Ergebnis von Verhaltensweisen. Schon Murphy und Kroeker (1988) kamen in ihrer Untersuchung bei der U.S. Navy in zu einem ähnlichen Schluss. Die Autoren weisen darauf hin, dass Job Performance eine zielgerichtete Aktivität ist. Deshalb müssen zuerst Ziele definiert werden, bevor daraus die gewünschte Job Performance abgeleitet werden kann. Die Job Performance fliesst dabei als abhängige Variable, das sogenannte Kriterium, in die Gleichung ein. Auf der anderen Seite stehen die über die Bewerbenden zur Verfügung stehenden Informationen als unabhängige Variablen, die sogenannten Prädiktoren, die das Kriterium vorhersagen sollen (Bortz & Schuster, 2010).

Die Prognosefähigkeit von SJTs wird häufig über Korrelationen mit ausgewählten Leistungsindikatoren angegeben (Kriteriumsvalidität). Unter der Kriteriumsvalidität wir der Grad des Zusammenhangs zwischen der betreffenden Skala (z.B. des SJT) mit nicht in der Skala erfassten Einstellungen oder Verhaltensmerkmalen (z.B. die Job Performance) verstanden. Je höher der Zusammenhang, desto besser ist die Kriteriumsvalidität zu bewerten (Reinecke, 2022). Ebenfalls kann die Beziehung von den SJT-Ergebnissen mit einer anderen Variable mithilfe von einer Faktoranalyse berechnet werden. Diese gibt darüber Auskunft, wie gut ein SJT-Item zur restlichen SJT-Itemgruppe passt im Hinblick auf den Zusammenhang mit der betreffenden Variable. Sowohl die Kriteriumsvalidität als auch die Faktoranalyse messen jedoch nur Zusammenhänge und können keine Prognose in eine Richtung abgeben. Möchte man Aussagen zur prädiktiven Validität eines Instruments machen, so wird häufig eine lineare oder multiple Regression gerechnet. Damit können

Änderungen in einer Kriteriumsvariablen (z.B. Job Performance) durch Änderungen in einer oder mehreren Prädiktorvariablen (z.B. SJT-Werten) vorhergesagt werden (Bortz & Schuster, 2010).

Studien zur Prognosefähigkeit von SJT in der Personalauswahl findet man überwiegend im akademischen und speziell im medizinischen Kontext. Dies hängt unter anderem damit zusammen, dass gerade im medizinischen Bereich auch nicht-akademische Fähigkeiten (soziale Kompetenzen) eine grosse Rolle spielen und zu deren Messung häufig SJTs verwendet werden. Der positive Zusammenhang zwischen sozialen Kompetenzen (Inventar sozialer Kompetenzen von Kanning, 2009) und der beruflichen Leistung konnte bereits von Jansen, Melchers und Kleinmann (2012) bestätigt werden. Sie konnten signifikante Zusammenhänge zwischen den sozialen Kompetenzen «Soziale Orientierung» und «Selbststeuerung» sowie dem Gesamtwert der sozialen Kompetenz mit der aufgabenbezogenen Leistung feststellen. Darüber hinaus konnten die Autoren einen substanziellen inkrementellen Beitrag der sozialen Kompetenz über die Ergebnisse von Assessment Centern hinaus zur Varianzaufklärung von aufgabenbezogener Arbeitsleistung zeigen. Lievens & Sackett (2012) zeigten in ihrer Studie mit 732 Medizinstudierenden einen signifikanten Zusammenhang zwischen den mit einem videobasierten SJT gemessenen sozialen Verhaltensweisen der Medizinstudierenden und der Job Performance aufgrund von Vorgesetztenbeurteilungen neun Jahre später. Webster et al. (2020) untersuchten in ihrem Review 26 vergleichbarer Studien zur Messung von nicht-akademischen Fähigkeiten von Bewerbenden für die medizinische Ausbildung. Die Ergebnisse zeigen, dass die SJT-Summenscores mindestens moderat mit anderen Instrumenten zur Leistungsbeurteilung zusammenhängen, die auf nicht-akademische, soziale Fähigkeiten zurückzuführen sind. Es zeigten sich durchschnittliche Zusammenhänge von $r = .32$ und ein Grossteil der Studien konnten statistisch signifikante Ergebnisse vorweisen ($p < .05$). Auch in einer früheren Metaanalyse von McDaniel et al. (2007) zur Prognosefähigkeit von SJT allgemein konnten schon vergleichbare Ergebnisse festgestellt werden. Die Autoren fanden eine

durchschnittliche Kriteriumsvalidität von r = .34. Die Job Performance wurde in den meisten der 39 berücksichtigten Studien wiederum in Form einer Vorgesetztenbeurteilung bewertet. Ausserdem konnte in der Metaanalyse gezeigt werden, dass bei SJT basierend auf Tätigkeitsanalysen im Vorfeld höhere kriteriumsbezogene Validitäten (r = .38) erreicht wurden als bei SJT ohne Tätigkeitsanalyse (r = .29).

Während SJT auch in anderen Branchen mit starkem sozialen Bezug wie beispielsweise im Ausbildungssektor (Bardach et al., 2021) oder in der Sicherheitsbranche (Leeds et al., 2003) in der Personalauswahl zum Einsatz kommen, greifen viele Autoren die Frage auf, ob video- oder textbasierte SJT bessere prädiktive Werte liefern. Die Ergebnisse dazu sind nicht immer ganz eindeutig. In der Metaanalyse von Christian, Edwards und Bradley (2010) wurde für videobasierte SJT eine bessere Kriteriumsvalidität gefunden als für textbasierte SJT. Vor allem für die Messung von interpersonale Kompetenzen waren die Ergebnisse besonders deutlich, da sich nicht einmal die Konfidenzintervalle überschnitten (videobasiert r = .47; textbasiert r = .27). Auch Lievens und Sackett (2006) konnten zeigen, dass videobasierte SJT eine signifikant höhere Kriteriumsvalidität aufweist als ein inhaltlich identischer Paper-Pencil-SJT. Ausserdem zeigte die textbasierte Version des SJT statistisch signifikant höhere Korrelationen mit der kognitiven Leistung der Befragten (Lievens & Sackett, 2006). Dem gegenüber stehen die Ergebnisse von Schäpers, Mussel, Lievens, König, Freudenstein und Krumm (2020). Die Autoren konnten zeigen, dass das Weglassen der Situationsbeschreibung bei text- und videobasierten SJT im gleichen Masse zu einer Verringerung des SJT-Scores führte und demnach die Stimuli den gleichen Einfluss die SJT-Scores haben. Ein möglicher Grund für diese uneinheitlichen Ergebnisse verschiedener Autorinnen und Autoren könnte sein, dass videobasierte SJT zusätzliche Varianz verursachen durch ethnische, geschlechtliche oder andere Faktoren wie das Alter der Schauspielenden. Individuen verarbeiten solche demografischen Informationen relativ schnell und vor allem automatisch, was das Antwortverhalten bei diesen Instrumenten beeinflussen könnte (Ito & Urland, 2003).

2.4 Zielsetzung und Fragestellungen

Aufbauend auf den beschriebenen theoretischen und empirischen Erkenntnissen, soll in der vorliegenden Masterarbeit das Ziel verfolgt werden, die Gelateria di Berna bei der jährlichen Rekrutierung von zahlreichen Verkaufsmitarbeitenden zu unterstützen. Aufgrund der schwierigen Situation auf dem Arbeitsmarkt soll der Fokus dabei auf der Attraction-Phase des Personalbeschaffungsprozess liegen. Um möglichst viele qualifizierte Stelleninteressierte anzusprechen und diese gleichzeitig bestmöglich bei ihrer Selbstselektion zu unterstützen, soll ein Online Self-Assessment entwickelt werden, das einen Einblick in den Berufsalltag von Verkaufsmitarbeitenden der GdB geben soll und wegen des hohen sozialen Bezugs im Gastgewerbe gleichzeitig die sozialen Kompetenzen der Befragten einschätzt. Um einen spielerischen Charakter zu erreichen und die Realitätsnähe des Instruments zu steigern, sollen die Videoaufnahmen mit einer 360°-Kamera gemacht und eine Hintergrundgeschichte simuliert werden. Ausserdem soll durch eine erste Testung des zu entwickelnden Instruments mit realen Bewerbenden der GdB dem vorherrschenden Forschungsdefizit im Bereich der Validität von digitalen Formaten in der Personalauswahl entgegengekommen werden. Um sowohl der praktischen als auch der empirischen Komponente der Zielsetzung nachzukommen, wird ein sequenzielles Mixed-Methods-Studiendesign angestrebt, bei dem in insgesamt drei Schritten ein videobasierter Situational Judgment Test als Grundlage für das Online Self-Assessment entwickelt und getestet werden soll. Aus dieser Zielsetzung werden folgende vier Leitfragen abgeleitet, die das methodische Vorgehen der Masterarbeit stützen sollen:

1. Welche Schlüsselszenarien aus dem Alltag von Verkaufsmitarbeitenden der Gelateria di Berna werden von Fachleuten als relevant bewertet und wie lauten möglich Verhaltensalternativen dazu?
2. Welche sozialen Kompetenzen sind besonders wichtig, um in diesen Schlüsselszenarien adäquat reagieren zu können?
3. Wie würden «ideale» Verkaufsmitarbeitende in diesen Schlüsselszenarien reagieren?

4. Wie ist die prädiktive Validität des 360°-videobasierten Situational Judgment Tests bezogen auf ausgewählte soziale Kompetenzen zu bewerten?

Im folgenden Kapitel soll das methodische Vorgehen zur Beantwortung dieser Leitfragen und zur Erreichung der geschilderten Zielsetzung im Detail erläutert werden.

3 Methodik

Nach der theoretischen und empirischen Einleitung soll in den folgenden Abschnitten die methodische Vorgehensweise zur Beantwortung der Leitfragen und damit zur Zielerreichung beschrieben werden. Es wird zuerst das Studiendesigns erläutert und begründet, bevor im weiteren Verlauf des Kapitels das Vorgehen im qualitativen und quantitativen Teil beschrieben wird. Dabei soll auf die Datenerhebung und -auswertung, die Zusammensetzung der Stichprobe sowie die eingesetzten Verfahren eingegangen und wo nötig mit Begründungen ergänzt werden.

3.1 Studiendesign

Die Zielsetzung der vorliegenden Masterarbeit besteht darin, ein digitales Instrument zur Selbstselektion (Online Self-Assessment) zu entwickeln und dessen prädiktive Validität bezüglich ausgewählter sozialer Kompetenzen zu testen. Da sich dieses Forschungsziel sowohl aus einer qualitativen als auch quantitativen Komponenten zusammensetzt, wurden für das Studiendesign qualitative und quantitative Methoden miteinander kombiniert zu einem sogenannten Mixed-Methods-Design. Die Kombination dieser unterschiedlichen Forschungsansätze begann erst im Laufe der 1980er-Jahre mehr und mehr an Akzeptanz zu gewinnen, ist heute aber ein etablierter Methodenansatz, der die Schwächen monomethodischer Ansätze ausgleichen soll (Baur, Kelle & Kuckartz, 2017). Mixed-Method-Designs sind klar von Multi-Method-Designs zu unterscheiden. Bei letzterem kommen zwar ebenfalls mehrere Methoden zum Einsatz, diese gehören aber jeweils der gleichen Methodenfamilie an (Kuckartz, 2014). Während quantitative Methoden mit standardisierten Erhebungsinstrumenten und numerischen Daten arbeiten, werden bei qualitativen Verfahren nicht-numerische Daten erhoben und es wird stärker auf die Sichtweisen, Einschätzungen und Motive der Forschungsteilnehmenden sowie deren Interaktion und Kommunikation mit den Forschenden fokussiert. Werden beide Methoden kombiniert, soll durch die Integration beider Stränge ein Mehrwert für die Forschung

entstehen (Greene, Caracelli & Graham, 1989). Die Methoden-Kombination der vorliegenden Masterarbeit ist wegen der aufbauenden Reihenfolge insbesondere dem sequenziellen Designtyp zuzuordnen (Qual → Quant). Dabei beeinflussen die Ergebnisse der vorgelagerten qualitativen Methode die Datenerhebung in der darauf folgenden quantitativen Methode. Ausserdem handelt es sich bei dieser Strategie insbesondere um ein exploratives Verallgemeinerungsdesign mit dem Ziel, ein Instrument zu entwickeln (Kuckartz, 2017).

In Abbildung 5 wird das geplante Studiendesign veranschaulicht. Die übergeordnete Zielsetzung soll in insgesamt drei Schritten erreicht werden. In dem vorgelagerten qualitativen Teil wird unter Berücksichtigung der Leitfragen 1 und 2 ein kompetenzbasierter Situational Judgment Test erarbeitet. Die qualitative Datenerhebung und -auswertung erfolgen dabei parallel während drei geplanter Workshops mit ausgewählten Fachleuten der GdB. In einem Zwischenschritt findet die Integration der beiden Forschungsansätze statt. Die qualitativen Ergebnisse werden dabei verfilmt und in eine Online Umfrage eingebettet. Im anschliessenden quantitativen Teil werden Daten mittels der erstellen Online Umfrage erhoben und anschliessend quantitativ ausgewertet. Die quantitativen Ergebnisse werden zur Beantwortung der Fragestellung 3 und 4 verwendet. Im weiteren Verlauf des Kapitels werden die einzelnen Schritte des geplanten Mixed-Methods-Design ausführlich beschrieben.

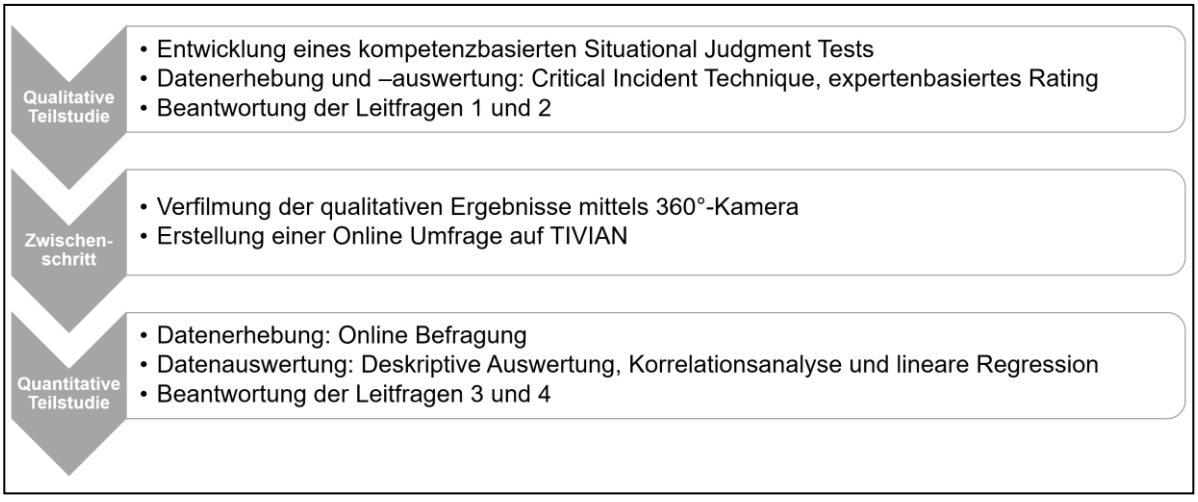

Abbildung 5. Grafische Darstellung des geplanten Studiendesigns (eigene Darstellung)

3.2 Qualitative Teilstudie

Übergeordnetes Ziel der qualitativen Teilstudie war, einen Situational Judgment Test zu entwickeln und damit die Inhalte für das angestrebte Online Self-Assessment zu erarbeiten. Dazu sollten relevante Schlüsselszenarien aus dem Alltag der Verkaufsmitarbeitenden der GdB gefunden (Leitfrage 1) und anschliessend kompetenzgeleitet ausgearbeitet werden (Leitfrage 2). Im Folgenden wird zuerst näher auf die qualitative Stichprobe eingegangen, bevor dann die Vorgehensweise und Wahl der verwendeten Methoden und Instrumente erläutert sowie begründet werden.

3.2.1 Stichprobe

Da es bei qualitativen Fragestellungen oftmals nicht um die Verteilung von Merkmalsausprägungen in einer Grundgesamtheit geht, spielt die Repräsentativität der Stichprobe für die Grundgesamtheit nur in seltenen Fällen eine Rolle. Vielmehr werden Stichproben bewusst so ausgewählt, dass der Erkenntnisgewinn maximiert werden kann (Schreier, 2020). Auch Kuckartz (2014) empfiehlt bei der Entwicklung eines Instruments mittels Verallgemeinerungsdesign (Qual → Quant) das sogenannte purposive Sampling (engl. für gezielte Fallauswahl). Aus diesem Grund wurde im qualitativen Teil der vorliegenden Masterarbeit die Stichprobenplanung an dessen Zielsetzung orientiert und es wurde eine im Voraus definierte gezielte Fallauswahl getroffen. Die Auswahl fiel auf acht Fachleute der GdB, die in ihrem Berufsalltag als Standortleitende der Filialen tätig sind. Als Fachleute (oder Experinnen und Experten) werden Personen bezeichnet, die bezogen auf das jeweilige Forschungsinteresse über spezifisches Wissen verfügen. Dabei wird dieses Wissen oftmals am entsprechenden Beruf festgemacht (Kühl, Strodtholz & Taffertshofer, 2009). So auch im vorliegenden Fall. In ihrer Position als Standortleitende sind die ausgewählten Fachleute nicht nur die Vorgesetzten der Verkaufsmitarbeitenden, sondern auch selbst regelmässig im Verkauf tätig. Dadurch wird ihr Prozess- und Deutungswissen als entsprechend hoch eingeschätzt. Gemäss Schreier (2020) ist diese Stichprobe als

homogen zu bewerten (gleichartige Fälle), was sich laut Autorin besonders gut eignet, um ein Phänomen im Detail zu beschreiben.

3.2.2 Critical Incident Technique

Die Critical Incident Technique (CIT) von Flanagan (1954) ist eine verhaltensbasierte, qualitative Methode, die von verschiedenen Autorinnen und Autoren zur Konstruktion von simulationsorientierten Verfahren wie dem Situational Judgement Test empfohlen (z.B. Kanning & Schuler, 2014; Muck, 2013; Weekley et al., 2015) und sollte deshalb auch zur Entwicklung des angestrebten SJT für die GdB verwendet werden. Die CIT besteht im Grundsatz aus drei Schritten. In einem ersten Schritt werden die Teilnehmenden nach erfolgskritischen Arbeitsereignissen gefragt. Auffällig ist, dass Unternehmen dabei oftmals nach negativen Ereignissen fragen, um zukünftige Fehler zu vermeiden. Die Frage nach spektakulären Erfolgen ist eher untypisch, aber hat einen sehr wertschätzenden Charakter (Serrat, 2017), der für den hier zu entwickelnden SJT genutzt werden sollte. In einem zweiten Schritt werden bei der CIT ähnliche Fälle gruppiert, ergänzt und diskutiert. Ziel dabei ist es, möglichst alle kritischen Ereignisse einer bestimmten Tätigkeit zu erfassen und sich über die auslösende Bedingung, die entsprechenden Verhaltensweisen und die Resultate daraus einig zu werden. Zuletzt werden die erarbeiteten Arbeitsereignisse nach ihrer Relevanz für den Tätigkeitserfolg bewertet (Blickle, 2019).

Ein grosser Vorteil der CIT ist, dass mit dieser Methode auch seltene Ereignisse aufgedeckt werden können, die bei einer Tätigkeitsanalyse eines klassischen Arbeitstages wie bei Christian et al. (2010) beschrieben vielleicht unentdeckt bleiben würden. Durch eine besondere Art des «Storytelling» können weitreichende Informationen zu der Motivation und den Emotionen von Personen gesammelt werden, was mit anderen Instrumenten schwieriger zu erfassen wäre (Serrat, 2017). Die Befragten bei der CIT erzählen aber aus einer sehr persönlichen, subjektiven Perspektive und es besteht die Gefahr, dass gewisse Informationen zurückgehalten werden, weil sie die erzählende Person selbst in schlechtes Licht rücken könnten. Ausserdem wird der CIT oft vorgehalten, dass meist nur wenige Fälle

zwar umfassend, aber nicht repräsentativ analysiert werden und dadurch keine repräsentative, generalisierbare Daten gewonnen werden können (Nixdorf, 2020).

3.2.3 Inventar sozialer Kompetenzen

Zur Einschätzung der Validität eines zu entwickelnden Instruments werden häufig bereits etablierte Verfahren in die Entwicklung miteinbezogen. Da laut einer Studie von Sisson und Adams (2013) im Gastgewerbe 86% der relevanten Kompetenzen von Mitarbeitenden zur erfolgreichen Berufsausführung soziale Kompetenzen sind, sollte für die kompetenzgeleitete Entwicklung des SJT das etablierte Inventar sozialer Kompetenzen (ISK) von Kanning (2009) hinzugezogen werden. Das ISK ist ein Selbstbeurteilungsinstrument in Form eines Fragebogens zur Erfassung zahlreicher allgemeiner sozialer Kompetenzen, die sowohl in privaten als auch in beruflichen Situationen von Bedeutung sind. «Allgemein» bringt dabei zum Ausdruck, dass die eingesetzten Skalen keinen Bezug zu einem bestimmten Beruf oder einem Anwendungsfeld haben. Vielmehr ist der ISK ein Breitbandverfahren, welches das Sozialverhalten von Menschen in unterschiedlichen Anwendungsfeldern wie beispielsweise Personalauswahl, klinischer Psychologie oder Pädagogik einschätzen kann. Für ein besseres Verständnis der zu messenden sozialen Kompetenz unterscheidet Kanning (2009) zwei Definitionen:

> **Sozial kompetentes Verhalten** = Verhalten einer Person, das in einer spezifischen Situation dazu beiträgt, die eigenen Ziele zu verwirklichen, wobei gleichzeitig die soziale Akzeptanz des Verhaltens gewahrt wird.
>
> **Soziale Kompetenz** = Gesamtheit des Wissens, der Fähigkeiten und Fertigkeiten einer Person, welche die Qualität des eigenen Sozialverhaltens – im Sinne der Definition von sozial kompetentem Verhalten – fördert.

Das sozial kompetente Verhalten ist dabei laut Autor immer situationsspezifisch, nicht aber die dem Verhalten zu Grunde liegende soziale Kompetenz. Da es sich bei der sozialen Kompetenz unstrittig um ein multidimensionales Konstrukt handelt, verweist Kanning (2009) auf die Notwendigkeit eines multidimensionales Verfahrens zu deren Messung. Der ISK soll

dieser Anforderung mit einer laut Autor überschaubaren Anzahl Items gerecht werden. Mit insgesamt 108 Items deckt der ISK 17 allgemeine soziale Kompetenzen (Primärfaktoren) ab, die sich faktoranalytisch zu vier abstrakteren Sekundärfaktoren gruppieren lassen: Soziale Orientierung, Offensivität, Selbststeuerung und Reflexibilität. Ebenfalls ist eine Kurzversion verfügbar mit nur 33 Items, die ausschliesslich Aussagen zu den Sekundärfaktoren erlaubt. In der Langversion des ISK werden pro Primärskala jeweils fünf bis neun Items in Form von verhaltensbasierten Aussagen präsentiert. Ein Beispielitem der Primärskala Prosozialität: «Auch wenn meine Zeit äusserst knapp bemessen ist, habe ich immer ein offenes Ohr für andere». Im Selbstbericht werden die Items auf einer vierstufigen Skala von *1 = trifft gar nicht zu* bis *4 = trifft sehr zu* bewertet. Die Durchführungsdauer bei der Langversion wird auf ca. 20 Minuten, bei der Kurzversion auf 10 Minuten geschätzt. Eine Übersicht der Primär- und Sekundärskalen des ISK findet sich in Tabelle 1.

Tabelle 1

Primär- und Sekundärskalen des ISK (Kanning, 2009, S. 28 ff.)

Sekundärskalen	Definition	Primärskalen
Soziale Orientierung	Ausmass, in dem eine Person anderen Menschen offen und mit positiver Grundhaltung gegenüber tritt	Prosozialität Perspektivenübernahme Wertepluralismus Kompromissbereitschaft Zuhören
Offensivität	Fähigkeit, aus sich herauszugehen und im Kontakt mit anderen Menschen eigene Interessen aktiv verwirklichen zu können	Durchsetzungsfähigkeit Konfliktbereitschaft Extraversion Entscheidungsfreudigkeit
Selbststeuerung	Fähigkeit eines Menschen, flexibel und rational zu handeln, wobei man sich selbst bewusst als Akteur begreift	Selbstkontrolle Emotionale Stabilität Handlungsflexibilität Internalität
Reflexibilität	Ausmass, in dem sich eine Person mit sich und ihren Interaktionspartnern aktiv auseinandersetzt	Selbstdarstellung Direkte Selbstaufmerksamkeit Indirekte Selbstaufmerksamkeit Personenwarhnehmung

Anmerkung. Die Definition bezieht sich auf die Sekundärskalen. Die Primärskalen sind als Bestandteile der Sekundärskalen zu verstehen.

Da der primäre Einsatzbereich des ISK im Personalsektor liegt, sind auch die Strichproben zur Normierung dementsprechend ausgerichtet. Nebst der zugrundeliegenden Gesamtstichprobe von 4208 Personen, stehen auch kleinere Teilstichproben daraus zur Verfügung: Schülerinnen und Schüler und Auszubildende, Studierende oder Berufstätige. Zur Auswertung steht ein vollständig standardisierter Auswertungsbogen zur Verfügung. Jedem Item wird je nach Antwort der Befragten ein Zahlenwert von 1 bis 4 zugeordnet, wobei bei negativ formulierten Items (ca. die Hälfte der Items) eine Umpolung vorgenommen werden muss. Pro Primärskala wird der Summerwert berechnet, woraus wiederum der Summerwert pro Sekundärskala summiert werden kann. Die Summenwerte bilden die Grundlage für die Interpretation anhand der gewünschten Normierungsstichprobe. Es sind je nach Fragestellung und Anwendungsbereich Aussagen zu Stanine, Prozentrang und Standardwerte möglich. Durch die ausführliche, schriftliche Instruktion und die standardisierten Auswertungs- sowie Interpretationsbeispiele wird eine hohe Durchführungs-, Auswertungs- und Interpretationsobjektivität als gesichert angesehen. Die internen Konsistenzen der Primärskalen für die Gesamtstichprobe liegen zwischen $\alpha = .69$ bis .84, für die Split-Half-Reliabilität werden Werte zwischen $r = .68$ bis .84 angegeben (Kanning, 2009). Die Reliabilität des ISK ist damit insgesamt als befriedigend bis gut zu bewerten. Durch Zusammenhangsanalysen mit anderen ausgewählten Instrumenten werden kriterien- sowie konstruktbezogene Validität aufgezeigt, welche den Ansprüchen der personaldiagnostischen Praxis gerecht werden (Scherp, 2010). Ebenfalls konnte das ISK in einer Studie von Jansen et al. (2012) signifikant zur Vorhersage der beruflichen Leistung beitragen. Auch für die vorliegende Masterarbeit wurde der ISK als passend eingeschätzt, da das Inventar mit sehr allgemeinen Items für einen breiten Einsatzbereich in Frage kommt und dennoch mittels eines gewünschten Referenzprofils sehr stellenspezifische Einschätzungen damit möglich sind. Ausserdem lassen sich die verhaltensbasierten Items der ISK gut mit der verhaltensbasierten Critical Incident Technique von Flanagan (1954) in Verbindung bringen.

3.2.4 Durchführung

Mit dieser Ausgangslange erfolgte die qualitative Datenerhebung und -auswertung im Rahmen von drei aufeinander aufbauenden Workshops mit den beschriebenen acht Fachleuten der GdB. Die Workshopinhalte wurden am vorgeschlagenen Prozess zur Entwicklung eines SJT von Muck (2013) orientiert und jeder Workshop hatte dabei eine spezifische Fragestellung und ein konkretes Ergebnis. Für eine bessere Übersicht sind die Inhalte und eingesetzten Methoden der Workshops in Tabelle 2 zusammengefasst.

Tabelle 2

Übersicht der qualitativen Workshops

	Fragestellung	Methode(n)	Ergebnis
Workshop 1	Mit welchen relevanten Szenarien werden Verkaufsmitarbeitende konfrontiert und wie sehen mögliche Verhaltensalternativen dazu aus?	Critical Incident Technique, Fokusgruppe	Konkrete Szenarien inkl. möglicher Verhaltensalternativen
Workshop 2	Welche Primärskalen des ISK sind zur Einschätzung der Leistung von Verkaufsmitarbeitenden besonders relevant?	Expertenbasiertes Rating	Schlüsselkompeten-zen von Verkaufs-mitarbeitenden
Workshop 3	Welche Schlüsselkompetenz unterscheidet in den jeweiligen Szenarien wesentlich zwischen Bewerbenden mit passendem und weniger passendem Antwortverhalten?	Expertenbasiertes, kontinuierliches Rating	Kompetenzbasierte Szenarien inkl. Rating der Verhaltens-alternativen (erste Version des SJT)

Anmerkung. Der Ablauf der Workshops orientiert sich am Entwicklungsprozess von SJT von Muck (2013).

Workshop 1. Ziel des ersten Workshops war die Beantwortung der Leitfrage 1, also die Erarbeitung konkreter, relevanter Schlüsselszenarien aus dem Berufsalltag von Verkaufsmitarbeitenden der GdB sowie passenden Verhaltensalternativen dazu. Dazu wurde die Critical Incident Technique von Flanagan (1954) eingesetzt. Für die konkrete Umsetzung wurde eine moderierten Fokusgruppe durchgeführt. Laut Blickle (2019) werden erfolgskritische Arbeitsereignisse meist mittels strukturierter Interviews mit Fachleuten erhoben. Jedoch können auch Fokusgruppen, Surveys, Aufzeichnungen der Arbeitsleistungen oder Arbeitstagebücher zum Einsatz kommen (Serrat, 2017). Im vorliegenden Forschungsprojekt wurde zur Erhebung bewusst ein Austausch zwischen den Fachleuten innerhalb einer Fokusgruppe gewählt, um die Erinnerungen durch gegenseitige

Erzählungen noch mehr anzuregen und damit möglichst viele verschiedene Schlüsselszenarien zu finden. Eine Fokusgruppe ist ein moderiertes Diskussionsverfahren zur Datensammlung. Ähnlich wie bei qualitativen Einzelinterviews wird anhand eines Leitfadens sichergestellt, dass alle relevanten Aspekte angesprochen werden und eine Orientierungshilfe gegeben ist (Schulz, Mack & Renn, 2012). Der erstellte Leitfaden für Workshop 1 orientierte sich an dem von Koch (2010) vorgeschlagenen Tool 1 des Workshop-Leitfaden vom Task-Analysis-Tool (TATOO) und umfasste im Wesentlichen folgende 4 Fragestellungen:

Leitfragen für Workshop 1
1. Welches sind die Ziele im Tätigkeitsbereich von Verkaufsmitarbeitenden der GdB?
2. Was sind Aufgaben, die im Tätigkeitsbereich von Verkaufsmitarbeitenden der GdB anfallen?
3. Welche typischen und alltäglichen Arbeitssituationen haben Sie schon selbst erlebt oder beobachtet?
4. Welches konkrete Verhalten hat dazu geführt, dass die Situation gut oder weniger gut bewältigt wurde?

Das Task-Analysis-Tool wurde in mehreren Studien empirisch geprüft und bietet durch den standardisierten Leitfaden eine gewisse Systematik (Höft & Goerke, 2014), die bei der CIT oftmals vermisst wird (Nixdorf, 2020). Auch Flanagan (1954) weist auf die Notwendigkeit einer möglichst präzisen und spezifischen Instruktion für die Beschreibung der Arbeitssituationen im Rahmen der CIT hin.

Nachdem in Schritt 1 der CIT relevante Arbeitssituationen und mögliche Verhaltensalternativen dazu gefunden wurden, ging es in Schritt 2 und 3 der CIT darum, die Arbeitssituationen inhaltlich zu gruppieren und nach Relevanz zu bewerten. Dazu wurden zusätzlich folgende Kriterien berücksichtigt, die in der Literatur als mögliche Gefahren oder nötige Bedingungen bei der Entwicklung eines SJT allgemein besprochen wurden (a-d) oder spezifisch für die Zielsetzung der vorliegenden Masterarbeit relevant sind (e und f):

Kriterien zur Einordnung und Bewertung der gefundenen Arbeitssituationen:
• Ist die Situation wirklich relevant zur Zielerreichung? (siehe Murphy und Kroeker, 1988) • Ist spezifisches Firmenwissen nötig, um die passende Verhaltensalternative zu finden? (siehe McDaniel, Whetzel, & Nguyen, 2006) • Wird wirklich ein Dilemma beschrieben? (siehe Bledow & Frese, 2009) • Ist die passendste Verhaltensalternative zu offensichtlich? (siehe Hough und Paullin, 1994) • Wirkt die Stelle oder die GdB in der Situation attraktiv genug? (Diercks, 2021) • Ist die Situation videobasiert darstellbar?

Für das weitere Vorgehen wurden nur diejenigen Arbeitssituationen berücksichtigt, die allen sechs Kriterien in der Diskussion mit den Fachleuten standgehalten haben.

Workshop 2. Ziel des zweiten Workshop war es, die wichtigsten sozialen Kompetenzen für die Verkaufsmitarbeitenden der GdB zu eruieren (Leitfrage 2). Dazu sollten die Fachleute einzeln die lange Version des ISK durchgehen und all diejenigen Items markieren, die gemäss ihrer Einschätzung relevant sind zur Beurteilung des Leistungsverhalten (Job Performance) von Verkaufsmitarbeitenden. Dieses Vorgehen ist angelehnt an das expertenbasierte Rating von Items von Bergman et al. (2006). Dabei sollen Experten Items bewerten aufgrund eines bestimmten Kriteriums (hier die Job Performance). Ausserdem hatte dieses Vorgehen den Vorteil, dass die Fachleute nicht mit den abstrakt formulierten, branchenfremden Begriffen der Primär- und Sekundärskalen des ISK, sondern direkt mit den verhaltensbasierten Items arbeiten konnten. Die Ergebnisse wurden anschliessend in der Gruppe diskutiert. Es wurden nur diejenigen Items des ISK berücksichtigt, die alle Fachleute nach der Diskussion als relevant zur Leistungsbeurteilung von Verkaufsmitarbeitenden bewertet haben. Aufgrund der relevanten Items konnten dann die entsprechenden Primärskalen des ISK ermittelt und die Schlüsselkompetenzen von Verkaufsmitarbeitenden bestimmt werden.

Workshop 3. Im dritten Workshop des qualitativen Teils wurden die Ergebnisse des ersten und zweiten Workshops zusammengeführt. Die übergeordnete Frage dabei war, welche Schlüsselkompetenz aus Workshop 2 jeweils in den gefundenen Szenarien aus

Workshop 1 wesentlich zwischen Bewerbenden mit passendem und weniger passendem Antwortverhalten unterscheiden kann. Das Vorgehen dazu wurde von Garman et al. (2006) übernommen. Die Fachleute sollten zuerst einzeln für jedes Szenario aus Workshop 1 entscheiden, welche der in Workshop 2 gefundenen Schlüsselkompetenzen am relevantesten ist, um in dem Szenario eine passende Antwort zu geben. Falls mehrere Schlüsselkompetenzen pro Szenario in Frage kamen, sollte eine Rangordnung erstellt werden. Ziel war es, pro Szenario eine Primärskala des ISK eindeutig zuweisen zu können. Da Lievens (2000) davor warnt, dass insbesondere bei der Beurteilung durch Fachleute vorkommen kann, dass sich die Fachleute nicht einig werden, wurde im Rahmen des Workshops die Möglichkeit offen gelassen, die Szenarien und die dazugehörigen Verhaltensalternativen so umzuformulieren, dass schlussendlich alle Fachleute mit der Zuteilung einverstanden waren. Falls keine eindeutige Zuteilung möglich war, wurde wie von Whetzel und McDaniel (2009) vorgeschlagen keine Zuteilung vorgenommen.

Die Fachleute sollten ausserdem ganz zum Schluss für jedes Szenario die Verhaltensalternativen in eine Rangfolge bezüglich der Effektivität des beschriebenen Verhaltens in der entsprechenden Situation bringen, was dem kontinuierlichen Rating entspricht (Polyhart & McKenzie, 2011). Wenn sich die Fachleute bei einer Rangfolge nicht einig waren, dann wurde über die Unterschiede diskutiert und die Verhaltensalternativen so umformuliert, bis sich die Gruppe einig war. Wäre die Gruppe der Fachleute grösser gewesen oder hätte man die Anzahl der Items im Rahmen der Entwicklung reduzieren wollen, so hätte auch die Interrater-Übereinstimmung berechnet werden können und man hätte die Szenarien mit einem zu tiefen Wert ausgeschlossen (Lievens & Sackett, 2006). Da im vorliegenden Fall die Anzahl Items schon sehr klein war (9 Szenarien) und sich die Fachleute schnell einig wurden, war dieser Schritt nicht nötig. Dabei muss beachtet werden, dass Items, die leichter übereinstimmend von Fachleuten bewertet werden, auch eher mit sozialer Erwünschtheit zusammenhängen (Krokos, Meade, Cantwell, Pond & Wilson, 2004, zitiert nach Whetzel & McDaniel, 2009, S. 196). Da Muck (2013) deutlich davon abrät,

diesen Schritt von den gleichen Fachleuten durchführen zu lassen, die die Szenarien in Workshop 1 entwickelt haben, wurde die Zuteilung zum Schluss von einer Person der Geschäftsleitung und zwei Verkaufsmitarbeitenden der GdB verifiziert.

3.2.5 Form des SJT und Auswertungsschlüssel

Grundsätzlich wurde für den zu entwickelnden SJT die Would-Do-Instruktion gewählt, obwohl diese Variante verfälschungsanfälliger ist (Nguyen et al., 2005). Grund für diese Wahl war der Anspruch, ein möglichst realistisches Erlebnis zu schaffen, was wiederum dazu führen könnte, dass die Befragten die Testsituation hinter dem Tool vergessen und dadurch der Einfluss der sozialen Erwünschtheit verringert wird (Georgiou et al., 2019). Um den Programmierungsaufwand für einen ersten Prototypen des Online Self-Assessment möglichst gering zu halten, wurde ausserdem ein klassisches Single-Choice-Format mit 4 Verhaltensalternativen pro Item gewählt. Wäre nach der besten und schlechtesten Verhaltensalternative, einem Ranking oder Rating der Verhaltensalternativen gefragt worden, so hätte das ausserdem den spielerischen Charakter des Online Self-Assessments geschwächt.

Der Auswertungsschlüssel gibt Auskunft darüber, wie die Antworten der Befragten in die Berechnungen der anschliessenden Datenauswertung einfliessen. In der Literatur findet man verschiedene Möglichkeiten bezüglich der Punktevergabe bei SJT, die im Theorieteil schon näher erläutert wurden. McDaniel et al. (2006) empfehlen, jede Antwortalternative auf einer vierstufigen Rating-Skala von den Befragten bewerten zu lassen, sodass ein grösseres Antwortspektrum angeboten wird und die Befragten nicht zu einer Antwort gezwungen werden. Ausserdem wird man so dem Umstand gerecht, dass es bei SJT-üblichen Fragestellungen meist keine objektiv richtige Antwort gibt, sondern mehrere Antwortalternativen plausibel, aber unterschiedlich effektiv sind (Bergman et al., 2006). Im vorliegenden Fall würde es aufgrund der Fragestellung «Wie würdest du reagieren?» und aufgrund des angestrebten Game-Charakters des finalen Online Self-Assessments wenig Sinn machen, wenn jede Verhaltensalternative bewertet werden müsste. Deshalb sollten

die Befragten jeweils pro Szenario nur die für sie passendste Alternative markieren. Ein Nachteil der «Forced-choice» ist jedoch, dass sie zu kategorialen Daten führt, die bei den statistischen Auswertungen zu Herausforderungen führen können (Lievens, 2000). Da sich in der Punktebildung die Effektivität der Verhaltensalternative abbilden sollte, wurde für den Auswertungsschlüssel eine absteigende Punktevergabe gewählt. Die laut Fachleute effektivste Antwort bekam 4 Punkte, die zweiteffektivste 3 Punkte, die dritteffektivste 2 Punkte und die schlechteste noch 1 Punkt.

3.3 Zwischenschritt (Integration)

Zur Integration der qualitativen und quantitativen Teile des Mixed-Methods-Designs sollten in einem Zwischenschritt die qualitativen Ergebnisse in eine quantitative Befragung überführt werden. Da für die Zielsetzung der Masterarbeit ein 360°-videobasierter SJT mit Game-Elementen vorgesehen ist, musste zuerst der qualitativ erarbeitete SJT verfilmt werden. Die videobasierte Präsentation der Stimuli wurde der textbasierten Version nach Abwägung der Argumente von Schäpers et al. (2020) vorgezogen, obwohl die Autoren keine Unterschiede zwischen den beiden Varianten bezüglich dem erreichten SJT-Score finden konnten. Die Autoren empfehlen eine genaue Abwägung von den Kosten eines videobasierten SJT (mehr Aufwand, ähnliche Resultate) und dessen Nutzen (höhere Augenscheinvalidität, verbesserte Bewerberbeteiligung und höhere prädiktive Validität für interpersonale Kriterien). Im vorliegenden Fall haben insbesondere die verbesserte Bewerberbeteiligung und höhere prädiktive Validität als Nutzen überwogen. Jedoch musste aus Zeitkosten-Gründen auf die Empfehlung von Schäpers et al. (2020), Stimuli- und Response-Komponenten im gleichen Format zu kombinieren, verzichtet werden.

Für die Aufnahmen wurde eine 360°-Filmkamera verwendet. Der Vorteil von 360°-Videos gegenüber normalen Filmaufnahmen liegt darin, dass sich die betrachtende Person bei 360°-Aufnahmen via Mausklick einmal um die eigene Achse drehen kann und der Effekt entsteht, als wäre man selbst im verfilmten Szenario. Die Aufnahmen hätten auch in einer

Virtual Reality Brille abgespielt werden können (siehe Kapitel 2.3.2). Jedoch ist für den Verwendungszweck der Praxispartnerin die Anwendung einer Web-Lösung sinnvoller, da das Selbstselektionstool von allen Stelleninteressierten zeit- und ortsunabhängig genutzt werden können soll. Bei der Entwicklung der Web-Lösung des 360°-videobasierten SJT wurde besonderen Wert darauf gelegt, dass die Umfrage sowohl an einem Laptop, als auch auf Mobiltelefonen durchgeführt werden kann. Zwar zeigt eine Studie von King, Ryan, Kantorowitz, Grelle und Dainis (2015), dass sich Befragte während einer Testung auf einem Mobiltelefon in ihren Möglichkeiten eingeschränkter fühlten als auf der PC-Version. Jedoch konnten andere Autoren vergleichbare Resultate für beide Formate berichten (Brown & Grossenbacher, 2017). Bei der vorliegenden Masterarbeit war vor allem wichtig, möglichst viele Bewerbende für die Umfrage zu gewinnen, um aussagekräftige statistische Auswertungen durchführen zu können. Indem die die Szenarien zum Schluss in eine passende Reihenfolge gebracht wurden, konnte mit einer Hintergrundgeschichte ein kompletter Einsatz von der Einsatzanfrage, über mehrere Interaktionen mit Gästen bis hin zum Feierabendbier mit dem Team simuliert werden.

Die finale Umfrage bestand aus insgesamt drei Teilen und wurde auf der Experience-Management-Software TIVIAN (2022) erstellt. Im ersten Teil sollten die Befragten über den Verwendungszweck der Daten und die Anonymität aufgeklärt werden. Ausserdem wurden die üblichen demographischen Daten erhoben. In einem zweiten Schritt sollte der videobasierte SJT eingebaut werden. Die verfilmten Szenarien wurden aneinander gereiht und in der Echtzeit-Entwicklungsplattform Unity® (2022) so programmiert, dass nach jedem Szenario automatisch die vier möglichen Verhaltensalternativen erschienen. In Abbildung 6 sind zwei verschiedene Ausschnitte aus dem entwickelten Tool zu sehen. Links bestellt eine Kundin an der Vitrine, rechts wird die befragte Person aufgefordert, sich für eine Verhaltensalternative zu entscheiden. Falls gewünscht, konnte das Szenario mit dem Replay-Knopf nochmals abgespielt werden. Unabhängig von der gewählten Verhaltensalternative wurde dann automatisch das nächste Video abgespielt. Der

videobasierte SJT wurde als Link in die Umfrage auf TIVIAN (2022) integriert. Nach dem letzten Video wurden die Befragten aufgefordert, zur TIVIAN-Umfrage zurückzukehren und die Befragung weiterzubearbeiten.

Abbildung 6. Ausschnitte aus dem 360°-videobasierten SJT.

Im dritten Teil der Online Umfrage wurde die lange Version des ISK eingebaut. Diese Daten wurden ausschliesslich für die Instrument-Entwicklung im Rahmen der Masterarbeit benötigt. Für das angestrebte Online Self-Assessment ist nur der videobasierte SJT angedacht. Aufgrund inhaltlicher und Effizienzgründen wurde beim ISK-Fragebogen auf die vierte Sekundärskala (Reflexibilität) verzichtet. Zum einen wurden die sozialen Kompetenzen dieser Skala von den Fachleuten als nicht relevant für die Bewertung der Job Performance von Verkaufsmitarbeitenden bewertet. Zum anderen sollte die Bearbeitungsdauer der freiwilligen Umfrage möglichst kurz gehalten werden. Ohne die vierte Sekundärskala waren es immer noch 83 ISK-Items, die im Rahmen der freiwilligen Online Umfrage bearbeitet werden sollten.

3.4 Quantitative Teilstudie

Nachdem im Zwischenschritt die Online Umfrage als Format zur Integration der qualitativen und quantitativen Methoden erstellt wurde, soll in diesem Kapitel die quantitative Methodik erläutert werden. Es wird zuerst auf die Stichprobengewinnung und das Vorgehen bei der Datenerhebung eingegangen. Anschliessend wird die methodische Herangehensweise zur Datenauswertung erläutert.

3.4.1 Stichprobengewinnung und Datenerhebung

Die Online Befragung fand im Zeitraum vom 29. Juni bis 16. August 2022 statt. Da die Stichprobe ausschliesslich aus realen Bewerbenden der GdB bestehen sollte, war im Vorfeld nicht bekannt, an wie viele Personen die Umfrage verschickt werden kann. Als Zielgrösse wurden 50 Teilnehmende angestrebt. Bei der Stichprobe handelt es sich grundsätzlich um eine systematische Zufallsstichprobe. Diese wird gemäss Döring und Bortz (2016) darüber definiert, dass die gesamte Zielpopulation bekannt ist und ab einem definierten Anfangszeitpunkt jede n-te Einheit für die Befragung ausgewählt wird. Im vorliegenden Fall ist die Zielpopulation definiert als die Gesamtheit aller Bewerbenden der GdB für die Stelle als Verkaufsmitarbeitende in einem bestimmten Zeitraum. Ab dem Zeitpunkt der Befragung wurde jeder bewerbenden Person die Online Umfrage geschickt. Die Zufallsstichprobe ist zu unterscheiden von der Gelegenheitsstichprobe, bei der keine Definition der Zielpopulation vorliegt, sondern willkürlich gut erreichbare Personen angefragt werden, wie beispielsweise in der Fussgängerzone. Die Stichprobe wurde aber davon beeinflusst, was für Anreize (siehe weiter unten) gesetzt wurden. Es liegt also ein gewisser Grad an Selbstselektion vor, die laut Autorenschaft aber grundsätzlich bei allen freiwilligen empirischen Untersuchungen vorliegt (Döring & Bortz, 2016). Ein weiterer Einfluss auf die Stichprobe konnte ausserdem die Wahl der Stellenplattformen gehabt haben. Die Stellenausschreibung war auf der Webseite der Gelateria di Berna auffindbar, von wo sie von einigen gängigen Schweizer Jobplattformen unkontrolliert im Internet geteilt wurde. Zudem wurde die Stelle auf den kostenlosen Universitäts- und Studierendenplattformen in Bern, Zürich und Basel aufgeschaltet.

Bewerbungen für die Stelle als Verkaufsmitarbeitende werden bei der Gelateria di Berna standardmässig ausschliesslich direkt über die Webseite angenommen. Der Link zur Online Befragung war im Rahmen der Masterarbeit noch nicht öffentlich auf der Webseite zugänglich, wie es für das finale Online Self-Assessment angedacht ist, sondern wurde vom HR-Team per E-Mail mit der Eingangsbestätigung der Bewerbung verschickt. Dabei wurde

explizit darauf hingewiesen, dass die Befragung freiwillig sowie anonym ist und die Ergebnisse keinen Einfluss auf den weiteren Bewerbungsprozess haben. Als Anreiz wurde ein erster Einblick in den Berufsalltag von Verkaufsmitarbeitenden sowie ein Gutschein für eine Portion Gelato und ein Feedback zu den persönlichen Ergebnissen in Aussicht gestellt. Der Link wurde an sämtliche Bewerbende (N = 130) im Umfragezeitraum verschickt. Es gab lediglich zwei Ausschlusskriterien, bei denen Bewerbende nicht für die Online Befragung eingeladen wurden: Bewerbende mit sehr wenig (Level A oder B) oder keinen Deutschkenntnissen und Bewerbende, die schon einmal in der Gelateria di Berna gearbeitet haben.

Um für die späteren Auswertungen ein Referenzprofil der ISK-Skalen zu erhalten, sollten auch die Fachleute der GdB die Online Umfrage einmal so ausfüllen, wie sie «ideale» Verkaufsmitarbeitende ausfüllen würden (maximal performance). Dieses Vorgehen ist angelehnt an den Prototypenansatz. Dieser Ansatz nutzt die Idee, dass Menschen für jede Eigenschaft eine prototypische Vorstellung haben. Wenn man nun die prototypische Vorstellungen verschiedener Personen sammelt, kann dann in der Summe ein typisches Bild einer Person mit dieser Eigenschaft konstruiert werden (Bühner, 2011). Diese erhaltenen Werte wurden separat zu den Werten der Bewerbenden erfasst, um die Ergebnisse nicht zu verfälschen.

3.4.2 Datenauswertung

Da der SJT und der ISK auf unterschiedlichen Plattformen abgefragt wurden, mussten vor der Datenauswertung die zwei Datensätze zusammengeführt werden. Bei der Umfrage im TIVIAN (2022) wurde allen Befragten automatisch eine 4-stellige Nummer zugewiesen, die bei der Weiterleitung via Link zu Unity® (2022) übernommen wurde. So konnten die Antworten einer Person in beiden Datensätzen problemlos zugeordnet werden. Die gesamte quantitative Datenauswertung fand mit dem Programm SPSS Statistics Version 28.0.1.1 (IBM, 2021) statt. Nachdem der Datensatz ins Programm eingelesen wurde, wurde er in einem ersten Schritt auf Vollständigkeit überprüft. Da beim SJT Personen mit

unvollständigem Antwortverhalten ausgeschlossen werden mussten (technische Probleme) und bei der Umfrage vom ISK alle Fragen als Pflicht markiert wurden, gab es keine fehlende Werte. Ausserdem gab es ausser der letzten Frage «Bemerkungen» keine Fragen, bei denen es zu Tippfehlern oder unmöglichen Werten kommen konnte, die die Resultate verfälschen hätten können. Als zweiter Schritt wurde die Item-Umpolung vorgenommen. Beim SJT waren die Antworten gemäss der zufällig angezeigten Reihenfolge in den Videos mit den Werten 1 bis 4 versehen, die so umgepolt wurden, dass die effektivste Antwort gemäss Urteil der Fachleute den Wert 4 und die ineffektivste Antwort den Wert 1 bekommt. Beim ISK wurde die Umpolung aller negativ formulierten Items gemäss Auswertungs-Manual von Kanning (2009) vorgenommen. Zudem wurden zusätzlich folgende Items für spätere Auswertungen berechnet: Summenscore der einzelnen SJT Items sowie Summenscore pro Primär- und Sekundärskala.

Deskriptive Analyse des SJT auf Itemebene. Beim SJT interessierten insbesondere Minimum, Maximum, Mittelwerte und Standardabweichungen sowie die Itemschwierigkeiten. Die psychometrische Itemschwierigkeit definiert den Grad der Zustimmung zu einem Item in Schlüsselrichtung der betreffenden Skala. Wird einem Item häufig gemäss der Schlüsselrichtung der Skala zugestimmt, so fällt der Mittelwert dieses Items hoch aus und das Item weist eine hohe psychometrische Schwierigkeit auf. Laut Bühner (2011) wird eine breite Streuung der Schwierigkeitsindizes angestrebt, um Personen möglichst gut differenzieren zu können. Die Normalverteilung der Daten wurde mittels Kolmogorov-Smirnov-Test, visueller Einschätzung der Histogramme und über die Werte der Schiefe und Kurtosis bewertet. Ausserdem wurden Boxplots zur Analyse von Ausreissern hinzugezogen. Von einer Reliabilitätsanalyse auf Itemebene wurde abgesehen, da die Items laut dem Urteil der Fachleute bis auf Item 1 und 8 alle unterschiedliche Konstrukte (Primärskalen des ISK) messen und sich deshalb nicht zu Skalen zusammenfassen lassen.

Deskriptive Analyse des SJT auf Summenscore Ebene. Um den Einfluss von demografischen Variablen auf die Höhe des Summenscores ausschliessen zu können,

wurde eine deskriptive Analyse der SJT-Summenscores nach Geschlecht, Alter und Beschäftigung vorgenommen. Es wurden ausserdem t-Tests für unabhängige Stichproben gerechnet und dabei die Voraussetzungen der Normalverteilung und Varianz innerhalb der Gruppen bei der Interpretation der Ergebnisse berücksichtigt. Um Aussagen über die Validität und Reliabilität des SJT machen zu können, muss wiederum die spezielle Skalenstruktur des SJT berücksichtigt werden. Die SJT-Items wurden als sogenannte Single-Item-Skalen behandelt, da sie fast alle unterschiedliche soziale Kompetenzen messen. Der totale Summenscore des SJT konnte nicht als Gesamtskala berücksichtigt werden. Um eine Schätzung der Reliabilität eines Tests vorzunehmen, wird in der Regel die interne Konsistenz berechnet. Dies kann jedoch nur bei Multi-Item-Skalen angewendet werden. Bei Single-Item-Skalen empfehlen Beierlein, Koveleva, László, Kemper & Rammstedt (2014) die Retest-Methode (Stabilität). Mit zwei Testergebnissen zu verschiedenen Zeitpunkten kann dann eine Korrelation gerechnet werden. Innerhalb der vorliegenden Masterarbeit war eine zweite Messung aus zeitlichen Gründen aber auch aus Gründen der Anonymität nicht möglich. Um dennoch einen Eindruck über die Zusammenhänge der Items untereinander zu bekommen, wurden die Inter-Item-Korrelationen berechnet. Die Konstruktvalidität des SJT wurde aufgrund der Spearman-Korrelationsanalyse mit den entsprechenden ISK-Primärskalen bewertet. Konkret handelt es sich dabei um die konvergente Validität, die dann vorliegt, wenn Messungen eines Konstrukts mittels verschiedener Instrumente hoch miteinander korrelieren (Schermelleh-Engel & Schweizer, 2003).

Deskriptive Analyse des ISK. Für eine erste deskriptive Analyse der ISK Ergebnisse wurden Mittelwerte, Standardabweichungen sowie Minimum und Maximum der Summenscores mithilfe der Normierungstabellen von Kanning (2009) in Standardwerte umgerechnet. Dabei wurde die Normierungstabelle der Gesamtstichprobe (N = 4208) verwendet, da diese ebenfalls hauptsächlich aus Studierenden (N = 1692), Berufstätigen (N = 1684) und Schülerinnen und Schüller sowie Auszubildenden (N = 689) zusammengesetzt

ist. Nur ein kleiner Teil der Normstichprobe (N = 143) stellen nicht Berufstätige und Rentner dar, die in der vorliegenden Stichprobe nicht vorkamen. Ausserdem wurden die internen Konsistenzen (Cronbachs Alpha) berechnet. Das Cronbachs Alpha ist ein Mass zur Messung der Zuverlässigkeit einer Gesamtskala (Janssen & Laatz, 2017). Da alleine die Standardwerte der ISK-Skalen nichts über die Eignung von Bewerbenden aussagen (Kanning, 2009), wurden ausserdem die standardisierten Mittelwerte der Stichprobe den standardisierten Mittelwerten des Referenzprofils von idealen Verkaufsmitarbeitenden gegenübergestellt.

Korrelationsanalyse. Da zur Berechnung einer linearen Regression ein linearer Zusammenhang zwischen den abhängigen und unabhängigen Variablen vorausgesetzt wird (Bortz & Schuster, 2010), wurden zur Voranalyse die Zusammenhänge zwischen den SJT Items mit den von den Fachleuten zugeordneten Primärskalen des ISK gerechnet. Da alle Items von einer Normalverteilung abwichen, eine Ordinal-Skalierung vorlag und Ausreisser vorhanden waren, wurde das nicht-parametrische Korrelationsverfahren nach Spearman herangezogen. Der Rangkorrelationskoeffizient r_s ist ein Zusammenhangsmass für ordinalskalierte Variablen basierend auf dem Produkt-Moment-Korrelationskoeffizient r nach Pearson (Janssen & Laatz, 2017). In die Korrelationsanalyse wurden ausschliesslich die von den Fachleuten erwarteten Zusammenhänge aufgenommen. Aufgrund der Mehrfachvergleiche wurde eine Anpassung des Signifikanzniveaus mit der Bonferroni-Korrektur gerechnet (Armstrong, 2014).

Prädiktive Validität des SJT. Zur Beantwortung der Leitfrage 4 nach der prädiktiven Validität des SJT bezogen auf die sozialen Kompetenzen des ISK sollte eine Regressionsanalyse erfolgen. Da im qualitativen Teil den entwickelten SJT-Items unterschiedliche Primärskalen des ISK zugewiesen wurden, wurde die prädiktive Validität mit Single-Item-Skalen berechnet. Laut Diamantopoulos, Sarstedt, Fuchs, Wilczynski und Kaiser (2012) ist die Berechnung der prädiktiven Validität von Single-Item-Skalen möglich und unter bestimmten Voraussetzungen sogar vergleichbar mit der Einschätzung durch

Multi-Item-Skalen. Zur Analyse der prädiktiven Validität der einzelnen Items wurde das Vorgehen von Bergkvist und Rossiter (2007) übernommen, welches auch schon von anderen Autoren adaptiert wurde. Bergkvist und Rossiter (2007) empfehlen die Berechnung von einzelnen linearen Regressionen pro Item, da die Ergebnisse einfacher zu interpretieren und Vergleiche zwischen den Analysen und einfacher möglich sind. Insgesamt wurden fünf lineare Regressionen gerechnet, da für fünf SJT-Items eindeutige Zuweisungen von Primärskalen des ISK möglich waren. Da das Antwortformat im SJT lediglich ein ordinales Skalenniveau aufweist (keine einheitliche Abstände zwischen den einzelnen Antwortalternativen), mussten die Antworten dummy-codiert werden, um als Prädiktor in die lineare Regression einfliessen zu können, da ein intervallskaliertes Skalenniveau Voraussetzung für eine lineare Regression ist (Hedderich & Sachs, 2020). Um eine Ordinalskala dennoch als «pseudometrische Skala» behandeln zu können, müssten laut Baur (2011) mindestens fünf verschiedene Ausprägungen vorliegen, deshalb wurde dieses Vorgehen hier nicht angewendet. Als Referenzskala wurde jeweils die passendste Verhaltensalternative verwendet. Dies ist damit zu begründen, dass die Itemschwierigkeiten sehr hoch ausfielen und deshalb nur sehr wenige Personen pro Item die unpassendste Antwort gewählt haben. Würde die unpassendste Antwort als Referenzskala verwendet werden, könnte dies laut Urban und Mayerl (2018) zu Schätzproblemen führen. Als Kriteriumsvariable (abhängige Variable) wurde jeweils der Summenscore der zugewiesenen Primärskala des ISK in die Regressionsanalysen aufgenommen. Dieser ist als intervallskaliert zu bewerten.

Zur Voraussetzungsprüfung wurde für jede Regressionsanalyse einzeln die Normalverteilung der Residuen mittels Histogramm beurteilt. Um die Unabhängigkeit der Residuen auszuschliessen wurde der Durbin-Watson-Koeffizient geprüft. Die Voraussetzung gilt gemäss Faustregel von Schneider und Ruoff (2010) als erfüllt, wenn der Wert zwischen 1.5 und 2.5 liegt. Die Voraussetzung der Homoskedastizität wurde mittels Streudiagramme der z-standardisierten vorhergesagten Werte und den z-standardisierten

Residuen kontrolliert. Die Residuen sollten unsystematisch um den Nullpunkt schwanken (Bortz & Schuster, 2010). Um den Verdacht auf Multikollinearität ausschliessen zu können, wurde zudem überprüft, ob die VIF-Werte < 10 sind (Marquardt, 1970). Für Aussagen zur prädiktiven Validität mittels Einschlussmethode wurden folgende Werte berücksichtigt: der standardisierte Regressionskoeffizient (β), der Standardfehler des Schätzers (SE), das quadrierte Bestimmtheitsmass (R^2), das korrigierte quadrierte Bestimmtheitsmass (korr. R^2) und das Signifikanzniveau.

4 Ergebnisse

Im Folgenden werden die Ergebnisse der qualitativen und quantitativen Teilstudie präsentiert und kommentiert. Die Ergebnisse werden chronologisch gemäss des verwendeten sequenziellen Mixed-Methods-Designs aufgeführt. Eine detaillierte Auseinandersetzung und Interpretation der Ergebnisse folgt im anschliessenden Kapitel 5.

4.1 Qualitative Ergebnisse

In insgesamt drei Workshops wurde ein SJT mithilfe ausgewählter Fachleute der GdB erarbeitet. Im Folgenden wird zuerst auf die Zwischenergebnisse der einzelnen Workshops eingegangen und die Erarbeitung des SJT anhand von Beispielen veranschaulicht, bevor der finale SJT in voller Länge präsentiert wird.

4.1.1 Zwischenergebnisse aus den Workshops

In Workshop 1 wurden insgesamt 25 relevante Arbeitssituationen aus dem Alltag von Verkaufsmitarbeitenden der GdB gefunden, wovon unter Berücksichtigung der sechs Kriterien a-f schlussendlich neun Szenarien mit jeweils vier Verhaltensalternativen für die Weiterentwicklung eruiert wurden (Leitfrage 1). Tabelle 3 gibt eine grobe Übersicht über die Entscheidungsfindung zum Ausschluss der restlichen 16 Situationen in Workshop 1.

Tabelle 3

Ausschluss-Entscheidungen während der CIT

Kriterium	Anzahl Situationen ausgeschlossen	Beispiel - Situation
(a) Relevanz	0	-
(b) Spezifisches Firmenwissen	8	Gast möchte eine Sorte, die nicht in der Vitrine ist.
(c) Fehlendes Dilemma	2	Gast probiert eine Sorte und mag sie nicht.
(d) Schwierigkeit	2	Gast fragt nach einem Glas Wasser.
(e) Attraktivität	3	Fremdkörper wird im Gelato entdeckt.
(f) Video-Tauglichkeit	1	Schichtablösung erscheint nicht.

Anmerkung. Die Kriterien werden in der Methodik ausführlich beschrieben.

Die meisten Szenarien mussten ausgeschlossen werden, weil spezifisches Firmenwissen nötig gewesen wäre, um die Situation richtig einzuschätzen (Kriterium b). In zwei Szenarien war das Dilemma zu schwach ausgeprägt (Kriterium c), bei zwei anderen Szenarien konnten nicht genügend verschiedene Verhaltensalternativen gefunden werden, ohne dass die passendste Antwort zu offensichtlich gewesen wäre (Kriterium d). Drei Szenarien wurden aufgrund mangelnder Arbeitgeber-Attraktivität ausgeschlossen (Kriterium e), eine Situation war schwierig in einem Video darzustellen (Kriterium f).

Im zweiten Workshop wurden aufgrund der Bewertung der ISK-Items durch die Fachleute insgesamt sieben soziale Schlüsselkompetenzen von Verkaufsmitarbeitenden eruiert (Leitfrage 2): Prosozialität, Wertepluralismus, Extraversion, Entscheidungsfreudigkeit, Selbstkontrolle, Emotionale Stabilität und Handlungsflexibilität. In Tabelle 4 werden nebst den dazugehörigen Sekundärskalen auch die jeweils am häufigsten genannten Items dazu aufgeführt. Von der Sekundärskala Reflexibilität wurde keine Primärskala als Schlüsselkompetenz festgestellt.

Tabelle 4

Schlüsselkompetenzen von Verkaufsmitarbeitenden der GdB

Sekundärskala	**Primärskala**	**Beispiel-Item**
Soziale Orientierung	Prosozialität	«Auch wenn meine Zeit äusserst knapp bemessen ist, habe ich immer ein offenes Ohr für andere» (Item 85)
	Wertepluralismus	«In meinem Alltag habe ich besonders gerne mit den unterschiedlichsten Typen von Menschen zu tun» (Item 47)
Offensivität	Extraversion	«Ich gehe immer auf Menschen zu, wenn ich sie kennenlernen möchte» (73)
	Entscheidungsfreudigkeit	«Probleme nehme ich sofort in den Angriff, wenn sie entstehen» (82)
Selbststeuerung	Selbstkontrolle	«Ich lebe immer nach der Devise: «Erst denken, dann handeln.» (Item 31)
	Emotionale Stabilität	«Grundsätzlich bin ich nicht leicht aus der Ruhe zu bringen» (Item 44)
	Handlungsflexibilität	«Auch in ausweglosen Situationen weiss ich immer, wie ich mich am besten verhalte» (Item 49)
Reflexibilität	-	-

Anmerkung: Zur Sekundärskala Reflexibilität wurden keine Schlüsselkompetenzen gefunden.

In Workshop 3 wurden die neun Schlüsselszenarien wo möglich mit den Schlüsselkompetenzen aus Workshop 2 kombiniert. Für Szenario 1, 2, 3, 5 und 8 konnten sich die Fachleute eindeutig auf je eine Primärskala einigen, die zwischen Bewerbenden mit passendem und weniger passendem Antwortverhalten unterscheiden kann. Szenario 4, 6, 7 und 9 konnten entweder keiner oder nicht eindeutig einer Primärskala zugeordnet werden. In Tabelle 5 wird das Endergebnis des qualitativen Teils, der entwickelte Situational Judgment Test, inklusive der Zuordnung der sozialen Kompetenzen aufgeführt. Insgesamt haben sich die Fachleute auf neun relevante Szenarien (Items) aus dem Arbeitsalltag von Verkaufsmitarbeitenden geeinigt und wo möglich unter Berücksichtigung der jeweiligen Primärskalen passende Verhaltensalternativen dazu formuliert. In Klammer steht jeweils die Codierung (*4 = effektivste Antwort*, absteigend).

Tabelle 5

Endergebnis des qualitativen Teils: der Situational Judgment Test

	Items	Primärskala
Szenario 1:	Whatsapp-Chat: «Ciao a tutti – wir haben heute Nachmittag einen unerwarteten Ausfall. Könnte uns jemand spontan zwischen 14-18h hinter der Vitrine aushelfen?»	Prosozialität
Frage:	Du hast eine Whatsapp-Nachricht im Gruppenchat von deinem Team erhalten. Wie würdest du reagieren? a. Auch wenn ich sehr viel um die Ohren habe, werde ich alles versuchen, damit ich mein Team heute Nachmittag unterstützen kann. (4) b. Ich melde mich erst, wenn sich sonst niemand meldet. (3) c. Weil ich das Mindestpensum von 3 Schichten pro Woche einhalte, kann ich mit gutem Gewissen frei machen. (2) d. Gruppenchats sind einfach nur nervig. Erst mal auf stumm schalten! (1)	
Szenario 2:	Eine Kundin betritt den Laden und lässt sich nach der Begrüssung sehr viel Zeit mit der Begutachtung der Auswahl. Es vergeht einige Zeit und nichts passiert.	Extraversion
Frage:	Unsere Kundin tut sich offensichtlich sehr schwer mit der Auswahl und die Warteschlange hinter ihr wird immer länger... Wie würdest du reagieren? a. Ich gehe sofort auf die Kundin zu, biete ihr etwas zum Probieren an und versuche sie bestmöglich bei der Auswahl zu unterstützen. (4) b. Ich lasse die Kundin wissen, dass sie sich jederzeit einfach bei mir melden kann, wenn sie Hilfe braucht. Unterdessen bediene ich den nächsten Gast. (3) c. Die Kundin würde sich ja sicher melden, falls sie eine Frage hat. Es wäre mir unangenehm, sie zu hetzen. Deshalb warte ich aufmerksam und bin bereit. (2) d. Ich erledige in der Zwischenzeit andere Sachen, bis sich die Kundin entschieden hat. (1)	
Szenario 3:	Die gleiche Kundin stellt eine Frage: «Hat es im Gelato Nuss drin»?	Selbstkontrolle
Frage:	Die Kundin hat eine Frage zu den Inhaltsstoffen der Gelati. Die Warteschlange hinter ihr wird langsam ungeduldig. Wie würdest du reagieren? a. Ich nehme mir ausreichend Zeit, um die Information im Sorten-Lexikon nachzuschlagen, um wirklich ganz sicher zu sein. Dann müssen die anderen Gäste halt noch länger warten. (4) b. Ich würde eigentlich gerne vorwärts machen, weil die anderen Gäste schon lange warten. Ich rufe meiner Kollegin, damit sie übernimmt und ich endlich weiterbedienen kann. (3) c. Ich weise die Kundin darauf hin, dass gerade sehr viel los ist und wir deshalb diese Information nur nachschlagen können, wenn es wirklich um eine Allergie geht. (2) d. Ich bin sicher, dass es keine Nüsse im Cioccolato drin hat und kann deshalb sehr schnell weiterbedienen. (1)	
Szenario 4:	Nach langem Suchen bezahlt die gleiche Kundin mit vielen kleinen Münzen und verabschiedet sich.	
Frage:	Unser Gast hat endlich bezahlt und den Laden verlassen. Was würdest du als nächstes machen? a. Münzen versorgen, meine Hände gründlich waschen und desinfizieren. Anschliessend sofort weiterbedienen. (4) b. Münzen versorgen, meine Hände desinfizieren. Anschliessend sofort weiterbedienen. (3). c. Münzen versorgen, dann die Oberfläche und meine Hände kurz mit einem Putzlappen reinigen. Anschliessend sofort weiterbedienen. (2) d. Münzen versorgen, sofort den nächsten Gast bedienen, um keine Zeit mehr zu verlieren. (1)	
Szenario 5:	Eine neue Kundin betritt den Laden und stellt nach kurzer Zeit fest, dass ihre Lieblingssorte nicht verfügbar ist. Sie äussert ihre Unzufriedenheit sehr deutlich.	Emotionale Stabilität
Frage:	Die Kundin ist sehr unzufrieden und enttäuscht. Mango ist für heute aber definitiv ausverkauft, daran kannst du nichts ändern. Wie würdest du reagieren? a. Die Reaktion der Kundin ist sicher etwas übertrieben, aber ich versuche sie freundlich abzulenken und schlage ihr direkt eine Alternative vor. (4)	

Tabelle 5 (Fortsetzung)

	b. Ich muss einmal tief durchatmen, um selbst ruhig zu bleiben. Ich bediene die Kundin dann freundlich weiter und biete etwas anderes zum Probieren an. (3) c. Ehrlich gesagt weiss ich nicht, wie ich reagieren soll. Falls möglich hole ich mir Unterstützung von meinem Team, um in der Hektik nichts falsches zu sagen. (2) d. Auch wenn ich es nicht zeige, bin ich schon etwas wütend, schliesslich kann ich nichts dafür. Ich lasse meinen Ärger aber nicht an der Kundin aus. (1)	
Szenario 6:	Suchbild mit verschiedenen möglichen Aufgaben, um sich zu beschäftigen.	
Frage:	Es ist ruhig geworden in der Gelateria und aktuell gibt es keine Gäste zu bedienen. Wie würdest du reagieren? a. Zuerst muss unbedingt die Vitrine gereinigt werden und ich sammle ein paar Servietten vom Boden auf. So sieht es wieder ordentlich aus, wenn die nächsten Gäste kommen. (4) b. Die Cornetti müssen dringend aufgefüllt werden. Dann frage ich meine Kollegin in der Produktion, ob sie Hilfe braucht. (3) c. Zuerst frage ich kurz bei der Kollegin in der Produktion nach, ob sie Hilfe braucht. Danach kümmere ich mich um meinen Arbeitsbereich. (2) d. Besser jetzt Pause machen, sonst komme ich gar nicht mehr dazu. Sauber machen kann ich auch gleich noch! (1)	
Szenario 7:	Die Arbeitskollegin begrüsst neue Gäste und dabei fällt ihr etwas auf den Boden. Sie bittet darum, kurz zu helfen und bedient die Gäste weiter.	
Frage:	Deiner Arbeitskollegin ist in der Eile ein Putzlappen auf den Boden gefallen. Sie bittet dich, kurz auszuhelfen, da sie gerade die nächsten Gäste bedienen muss. Wie würdest du reagieren? a. Ich hebe den Putzlappen auf, werfe ihn in den Wäschekorb und lege einen frischen Putzlappen bereit. Dann helfe ich beim Bedienen der Gäste. (4) b. Ich hebe den Putzlappen schnellstmöglich auf (5-Sekundenregel) und lege ihn wieder bereit. Dann helfe ich beim Bedienen der Gäste. (3) c. Ich hebe den Putzlappen auf und wische gleich noch die Flecken auf der Vitrine weg. Dann werfe ich den Putzlappen in den Wäschekorb, lege einen frischen Putzlappen bereit und helfe beim Bedienen der Gäste. (2) d. Das Bedienen der Gäste hat immer oberste Priorität. Der Putzlappen muss warten. (1)	
Szenario 8:	Den gleichen Gästen fällt bei der Verabschiedung das Gelato auf den Boden.	Prosozialität
Frage:	Oh no! Unseren Gästen ist das Gelato auf den Boden gefallen. Wie würdest du reagieren? a. Ich mache den Gästen sofort kostenlos ein neues Gelato und wische dann alles weg. (4) b. Ich biete den Gästen einen frischen Putzlappen an und überprüfe, ob die restlichen Cornetti ebenfalls brüchig sind, damit den nächsten Gästen nicht das gleiche passiert. (3) c. Ich biete natürlich an, ein neues Gelato zu machen. Dieses müsste ich allerdings einkassieren, weil es nicht unser Fehler war. (2) d. Ich tue so, als hätte ich es nicht gesehen, damit es für die Gäste nicht peinlich ist. (1)	
Szenario 9:	Nach der Schicht wartet das Team draussen mit einer Runde Feierabendbier.	
Frage:	Dein Team lädt dich als Dank fürs Einspringen zum gemeinsamen Feierabendbier ein. Wie würdest du reagieren? a. Klar, sehr gerne! Das ist eine super Chance, alle besser kennenzulernen! b. Zuerst einmal brauche ich ein Gelato! c. Gibt's auch Rivella? d. Prost!	

Anmerkungen. In Klammer steht die Codierung für den Auswertungsschlüssel (4 = beste Antwort, absteigend).

4.2 Quantitative Ergebnisse

Zur Präsentation der quantitativen Ergebnisse der vorliegenden Masterarbeit wird zuerst die Stichprobe näher beschrieben. Im Anschluss werden die deskriptiven Ergebnisse des SJT und ISK einzeln betrachtet, bevor dann näher auf die Zusammenhänge der beiden Instrumente und damit die Validität des Instruments eigegangen wird.

4.2.1 Stichprobe

Insgesamt haben 102 Personen die Online Umfrage begonnen, wovon 57 Personen die Befragung komplett abgeschlossen haben. Das entspricht einer Beendigungsquote von 55.88%. Die meisten Abbrüche (34 von 45, also 75%) sind dabei auf der Seite aufgetreten, bei der die Weiterleitung via Link zum SJT aufgeführt war. Von den 57 abgeschlossenen Befragungsergebnissen auf TIVIAN (2022) mussten 6 Datensätze wegen fehlender Werte im SJT von der Auswertung ausgeschlossen werden. Es handelte sich dabei eindeutig um technische Fehler, da die fehlenden Werte unsystematisch waren, also keine Abbrüche, sondern Lücken im Datenset aufwiesen, obwohl Items nicht übersprungen werden konnten. Ausserdem hat eine Person die Befragung zweimal durchgeführt. In diesem Fall wurde nur der erste Durchlauf berücksichtigt. Insgesamt konnten 50 vollständige Datensätze in die Auswertung aufgenommen werden. Die durchschnittliche Bearbeitungszeit der Befragung lag bei 24 Minuten. Tabelle 6 gibt einen Überblick über die demografischen Merkmale der Strichprobe. Mit einem Anteil von 78% stellten Frauen eine grosse Mehrheit dar. Über die Hälfte der Personen (52%) waren ausserdem jünger als 23 Jahre. Das Mindestalter für eine Bewerbung bei der GdB und damit auch für die Teilnahme an der Online Umfrage war 18 Jahre. Gegen oben wurde keine Altersgrenze gesetzt, jedoch nahm die Anzahl Personen mit zunehmendem Alter erkennbar ab. Im Durchschnitt lag das Alter bei 23.21 Jahren (SD = 4.06), der Median liegt bei 22 Jahren. 68% der Bewerbenden befanden sich zum Zeitpunkt der Befragung im Studium, 22% waren hauptsächlich berufstätig, alle anderen gingen noch zur Schule (10%).

Tabelle 6

Demografische Merkmale der Stichprobe

		N	%	Mittelwert	SD	Median
Geschlecht	Weiblich	39	78			
	Männlich	6	12			
	Divers	3	6			
	Keine Angabe	2	4			
Alter	Total	50	100	23.21	4.06	22
	18-22	26	52			
	23-26	15	30			
	27 und älter	8	16			
	Keine Angabe	1	2			
Beschäftigung	Schüler*innen	5	10			
	Studierende	34	68			
	Berufstätige	11	22			

Anmerkung. Total N = 50. SD = Standardabweichung.

4.2.2 Deskriptive Analyse des SJT

In Tabelle 7 ist die Verteilung der SJT-Items abgebildet. Szenario 9 wird nur vollständigkeitshalber aufgelistet, aber ansonsten aus der Analyse ausgeschlossen, da dieses Item ausschliesslich für die Hintergrundgeschichte sowie zur Attraktivität des Stellenprofils im SJT berücksichtigt wird und es keine passende oder weniger passende Verhaltensalternativen gab. Die Item-Mittelwerte der restlichen Items bewegten sich zwischen 2.68 und 3.70 von maximal vier erreichbaren Punkten je Item. Die Standardabweichung nahm Werte zwischen SD = .087 und .166 an. Bis auf Item 3 wiesen alle Items relativ hohe Mittelwerte (> 3) auf, was auf hohe psychometrische Itemschwierigkeiten hindeutet (Bühner, 2011). Sowohl die visuelle Ansicht der Histogramme als auch die numerischen Daten zeigten eine rechtssteile Verteilung der Items (Schiefe < 0). Aufgrund der Kurtosis-Werte war festzustellen, dass einige Items stark von einer Normalverteilung abwichen (Kurtosis bei einer Normalverteilung = 0) und eine eher schmalgipflige Verteilung der Items vorlag (Kuhlmei, 2020). Auch der Kolmogorov-Smirnov-Test zeigte für alle Items signifikante Werte (< .001) an, was die Abweichung von der Normalverteilung nochmals bestätigte. Items 1, 4 und 8 schöpften nicht die komplette Bandbreite der Antwortmöglichkeiten aus, da jeweils keine Person die Antwort mit der

Codierung 1 gewählt hat. Die Analyse der Boxplots zeigte für die Items 1, 2, 5 und 6 jeweils zwischen 1 - 6 Ausreisser an, die unter den durchschnittlichen Werten lagen. Es wurden jedoch keine Datensätze ausgeschlossen, da insbesondere von der Norm abweichende Daten zum Erkenntnisgewinn beitragen können (Muck, 2013).

Tabelle 7

Itemanalyse und -kennwerte

	N	Min	Max	M	SD	Schiefe	Kurtosis
Szenario 1_PS	**50**	**2**	**4**	**3.70**	**.087**	**-1.92**	**2.54**
Szenario 2_EX	**50**	**1**	**4**	**3.46**	**.096**	**-1.29**	**2.20**
Szenario 3_SK	**50**	**1**	**4**	**2.68**	**.135**	**-.47**	**-0.63**
Szenario 4_Hyg1	50	2	4	3.28	.103	-.49	-.95
Szenario 5_ES	**50**	**1**	**4**	**3.44**	**.111**	**-1.50**	**2.06**
Szenario 6_Hyg2	50	1	4	3.46	.108	-1.30	1.09
Szenario 7_Hyg3	50	1	4	3.28	.159	-1.12	-0.42
Szenario 8_PS	**50**	**2**	**4**	**3.68**	**.101**	**-1.91**	**1.87**
Szenario 9_Bier	50	1	4	3.01	.166	-1.42	0.24

Anmerkungen. Fett gedruckt sind die kompetenzbasierten Items. PS = Prosozialität; EX = Extraversion; SK = Selbstkontrolle; Hyg1 = Hygiene 1; ES = Emotionale Stabilität; Hyg2 = Hygiene 2; Hyg3 = Hygiene 3; Bier = Szene Feierabendbier. M = Mittelwert. SD = Standardabweichung. Szenario 9 ist lediglich vollständigkeitshalber aufgelistet. Inhaltlich sind die Werte nicht interpretierbar.

Die Inter-Item-Korrelationen des SJT sind in Tabelle 8 dargestellt. Gemäss Kuckartz, Rädiker, Ebert und Schehl (2013) spricht man bei Werten von $.10 \leq |r| < .30$ von einem geringen Zusammenhang, was auf die meisten vorliegenden Korrelationen zutrifft. Bei tieferen Werten ist kein Zusammenhang mehr feststellbar. Bei Werten zwischen $|r| = .03$ bis .05 ist von einem mittleren Zusammenhang die Rede. Wäre das Ziel, eine Skala zu einem bestimmten Konstrukt zu bilden, so wären hohe mittlere Interkorrelationen zwischen den Items anzustreben (Kuckartz et al., 2013). Die höchste Korrelation war zwischen Item 4 und Item 8 festzustellen ($r = .33$). Die häufigste Ursache für negative Werte sind laut Blanz (2015) entweder, wenn Items vergessen wurden umzupolen oder wenn es sich um Items handelt, die etwas Gegensätzliches messen. In Tabelle 9 ist eine Übersicht über die Summenscores des SJT nach Geschlecht, Alter und Beschäftigung zu finden. Insgesamt konnten die Bewerbenden im SJT einen Summenscore von mindestens 8 bis maximal 32

erreichen (ohne Item 9). Der Mittelwert des Summenscores über die gesamte Stichprobe lag bei M = 26.98 (SD = 2.70, Min = 18, Max = 31).

Tabelle 8

Inter-Item-Korrelationsmatrix des SJT

Item	1	2	3	4	5	6	7	8
Szenario 1_PS	-							
Szenario 2_EX	**.24**	-						
Szenario 3_SK	.01	-.18	-					
Szenario 4_Hyg1	.01	.19	-.02	-				
Szenario 5_ES	-.14	-.20	.19	.03	-			
Szenario 6_Hyg2	-.14	.02	**.29**	**.28**	**.23**	-		
Szenario 7_Hyg3	-.11	-.20	**.27**	.15	.16	.09	-	
Szenario 8_PS	-.08	.10	-.06	**.33**	-.14	-.02	.06	-

Anmerkungen: Fett gedruckt sind die höchsten Korrelationen. PS = Prosozialität; EX = Extraversion; SK = Selbstkontrolle; Hyg1 = Hygiene 1; ES = Emotionale Stabilität; Hyg2 = Hygiene 2; Hyg3 = Hygiene 3. Szenario 9 wurde aus der Analyse ausgeschlossen.

Tabelle 9

Deskriptive Statistik SJT Summenscore

	N	Mittelwert	SD	Min	Max	Schiefe	Kurtosis
Gesamt-Stichprobe	50	26.98	2.70	18	31	-0.89	1.10
Geschlecht							
Weiblich	39	27.3	2.37	22	31	-0.50	-0.35
Männlich	6	25.8	4.62	18	30	-1.12	0.39
Divers	3	27.3	2.52	25	30	0.59	-
Keine Angabe	2	24.5	.71	24	25	-	-
Alter							
18-22	26	26.6	2.89	18	31	-1.09	1.83
23-26	15	27.3	2.41	23	31	-0.42	-0.13
27 und älter	8	27.1	2.75	23	30	-0.50	1.52
Keine Angabe	1	30.0	.00	30	30	-	-
Beschäftigung							
Schüler*innen	5	26.4	1.14	25	28	0.41	-0.18
Studierende	34	26.7	2.97	18	31	-0.86	0.69
Berufstätige	11	28.0	2.14	24	31	-0.60	-0.23

Anmerkung. N = 50. SD = Standardfehler.

Der Kolmogorov-Smirnov-Test wurde signifikant (p = .001), was gegen eine Normalverteilung spricht. Die Person mit dem kleinsten Summenscore wurde in den Boxplots als Ausreisser der Gesamtstichprobe ausgewiesen, wurde aber zu Gunsten des Erkenntnisgewinns nicht von den Auswertungen ausgeschlossen. Ausserdem war die Normalverteilung auch nach Ausschluss dieses Ausreissers nicht gegeben. In Abbildung 7 ist dazu die Verteilung des SJT Summenscores der Gesamtstichprobe inkl. Normalverteilungskurve dargestellt (Abszisse = Häufigkeit, Ordinate = Summenscore).

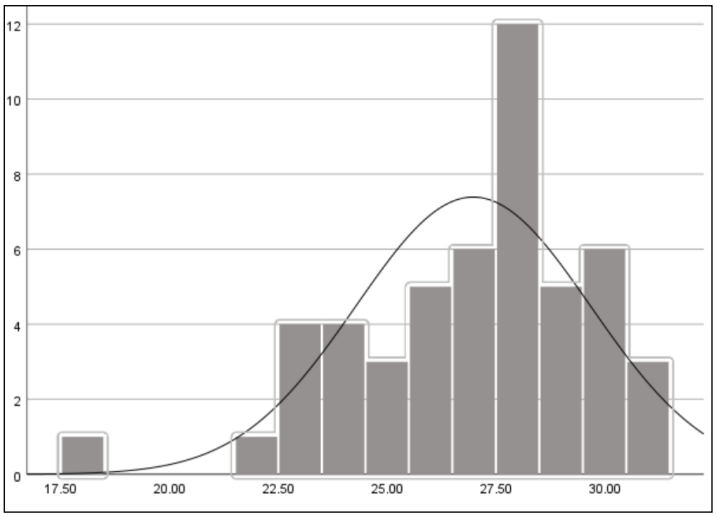

Abbildung 7. Verteilung des SJT-Summenscores der Gesamtstichprobe.

In den t-Tests wurden keine signifikanten Gruppenunterschiede für das Geschlecht, das Alter und die Beschäftigung gefunden, wobei zu beachten ist, dass die Voraussetzungen für einen t-Test unabhängiger Stichproben nicht immer gegeben waren. Da die Gruppengrössen sehr unterschiedlich waren (z.B. 29 Frauen vs. 6 Männer), hatten die signifikant abweichenden Varianzen einen Einfluss auf den p-Wert, was die Interpretation verfälschen kann. Ebenfalls war in den meisten Fällen keine Normalverteilung in den Variablen gegeben, was ebenfalls eine Voraussetzung für die t-Tests wäre. Für die weiteren Analysen wurde trotz suboptimalen Voraussetzungen davon ausgegangen, dass keine Gruppenunterschiede bezüglich der SJT-Werte vorlag.

Zum Schluss der Umfrage sollten die Bewerbenden ihre Erfahrungen mit dem SJT bewerten auf einer Likert-Skala von *1 = sehr schlecht* bis *5 = sehr gut*. Es wurde ein

Mittelwert von M = 4,3 (Min = 3, Max = 5, SD = .707) erreicht. Verbesserungsvorschläge und Kommentare der Bewerbenden werden in der Diskussion in Kapitel 5 näher erläutert.

4.2.3 Deskriptive Analyse ISK

Die sozialen Kompetenzen der Bewerbenden der GdB lagen verglichen mit der Normierungsstichprobe vom ISK im Durchschnitt, wobei die einzelnen Werte von weit unterdurchschnittlich (M < 80) bis weit überdurchschnittlich (M > 120) reichen. Die Primärskalen Zuhören (p ≥ .06), Konfliktbereitschaft (p ≥ .16), Extraversion (p ≥ .19) sowie Entscheidungsfreudigkeit (p ≥ .13) wiesen eine Normalverteilung auf. Alle anderen Primärskalen sowie die alle drei Sekundärskalen waren nicht normalverteilt (p ≤ .05). Die internen Konsistenzen lagen zwischen α = .42 und .85 bei den Primärskalen und α = .82 bis .88 bei den Sekundärskalen. Laut Blanz (2015) sind von .80 und höher als gut und sehr gut zu bewerten. Ist das Cronbach's Alpha niedriger als .50, spricht das für eine schlechte interne Konsistenz der Skala. Dies war jedoch nur bei der Primärskala Handlungsflexibilität der Fall (α = .42). In Tabelle 10 findet sich eine Übersicht der Skalenwerte des ISK.

Tabelle 10

Übersicht Skalenwerte ISK

	M	SD	Min	Max	α
Soziale Orientierung	**109.08**	**8.46**	**81**	**126**	**.88**
Prosozialität	106.10	7.89	78	121	.61
Perspektivenübernahme	107.60	10.17	84	120	.82
Wertepluralismus	108.92	7.00	94	124	.54
Kompromissbereitschaft	105.64	8.27	84	124	.61
Zuhören	105.78	8.02	76	120	.71
Offensivität	**100.46**	**8.46**	**80**	**118**	**.83**
Durchsetzungsfähigkeit	100.60	8.12	83	118	.64
Konfliktbereitschaft	98.10	7.41	85	114	.56
Extraversion	102.76	8.05	85	118	.78
Entscheidungsfreudigkeit	100.66	9.92	71	118	.85
Selbststeuerung	**106.52**	**7.79**	**81**	**120**	**.82**
Selbstkontrolle	107.52	6.70	92	121	.52
Emotionale Stabilität	104.82	6.79	82	121	.54
Handlungsflexibilität	104.26	6.78	90	119	.42
Internalität	102.10	9.22	80	121	.70

Anmerkung. N = 50. Es werden die T-Werte angegeben. Normierung mit der Gesamtstichprobe des ISK. Fett gedruckt sind die Werte der Sekundärskalen. M = Mittelwert. SD = Standardabweichung. α = Cronbachs Alpha.

In Abbildung 8 wurden die Standardwerte der ISK-Skalen der befragten Bewerbenden den Werten von «idealen» Verkaufsmitarbeitenden (Referenzprofil) gemäss Einschätzung der Fachleute gegenübergestellt. In fast allen Primär- und Sekundärskalen liegt der ideale Wert für das Jobprofil deutlich über dem Wert der Stichprobe. Eine Ausnahme bilden die Primärskalen Durchsetzungsfähigkeit, Konfliktbereitschaft und Kompromissbereitschaft. Auf der Primärskala Kompromissbereitschaft sind die Werte beider Gruppen fast identisch. Wie die Skalen zueinander im Verhältnis stehen, also die Form der beiden Profile, ist relativ ähnlich, jedoch in einem anderen Wertebereich.

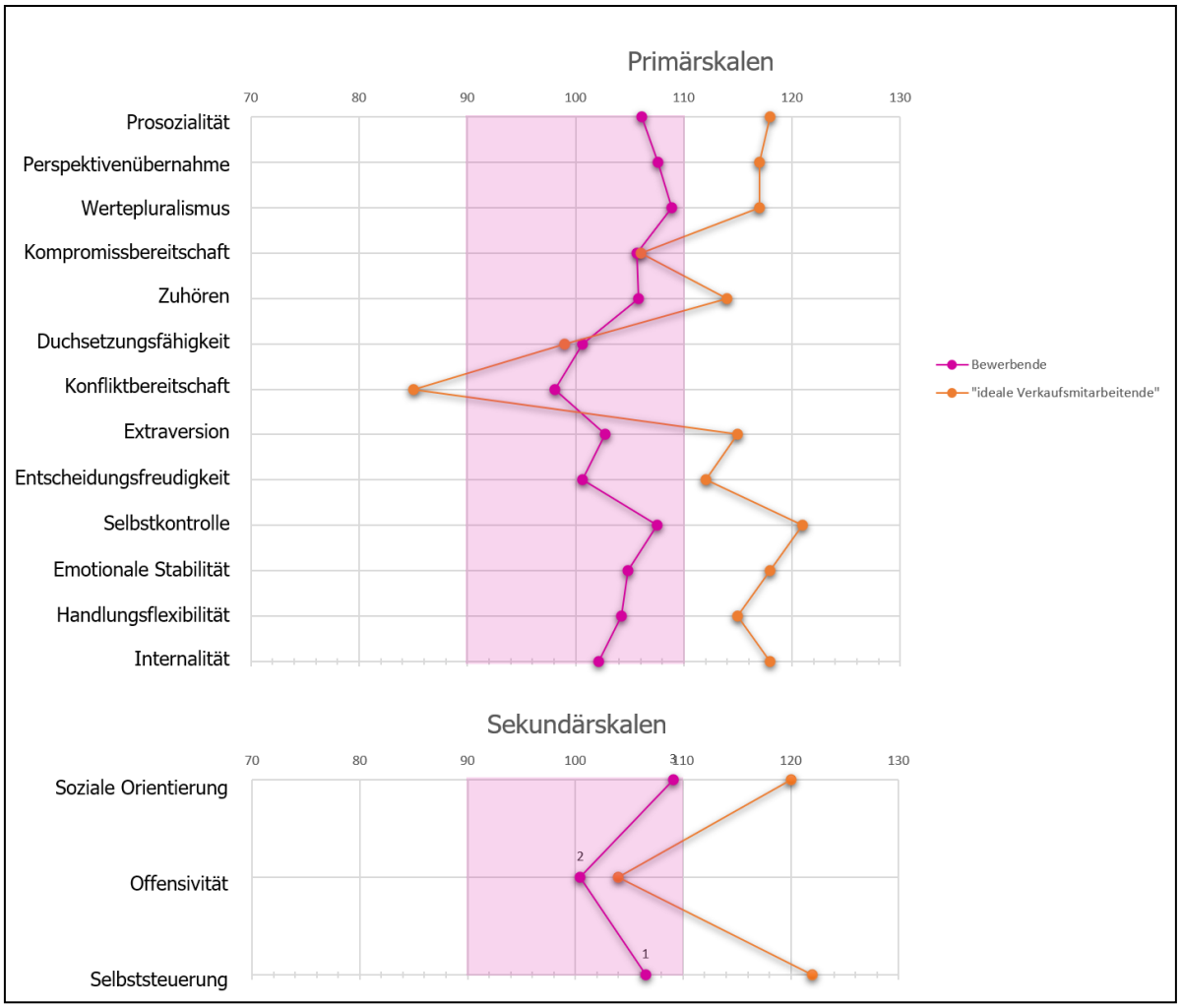

Abbildung 8. Referenzprofil von «idealen Verkaufsmitarbeitenden» und Standardwerte der Stichprobe.

4.2.4 Korrelationsanalyse SJT und ISK

In der qualitativen Teilstudie haben die Fachleute bei insgesamt fünf Szenarien eindeutig jeweils eine Primärskala des ISK zuweisen können. In Tabelle 11 werden für diese fünf Paare die jeweiligen Rangkorrelationen nach Spearman abgebildet. Bis auf Szenario 2 korrelieren alle SJT-Items positiv mit den jeweils zugeordneten ISK-Primärskalen. Das heisst, je besser die Antwort im SJT, desto höher die Werte auf der entsprechenden ISK-Skala. Der stärkste Zusammenhang wurde für Szenario 8 gefunden, welches eine Korrelation mit der Prosozialitätsskala in Höhe von $r_s = .21$ aufweist. Gemäss Kuckartz et al. (2013) entspricht dies einem geringen Zusammenhang. Auch für Szenario 5 und die Primärskala Emotionale Stabilität konnte ein geringer Zusammenhang gefunden werden ($r_s = .14$). Jedoch wurde schon vor der Bonferroni-Korrektur keiner der beiden Zusammenhänge signifikant. Für Szenario 1, 2 und 3 sind die Zusammenhänge fast null. Da nur sehr kleine Zusammenhänge gefunden wurden, ist auch die Konstruktvalidität (konvergente Validität) des SJT dementsprechend niedrig zu bewerten (Schermelleh-Engel & Schweizer, 2003).

Tabelle 11

Rangkorrelationen nach Spearman

	Prosozialität PS		Extraversion EX		Selbstkontrolle SK		Emotionale Stabilität ES	
	r_s	Sig.	r_s	Sig.	r_s	Sig.	r_s	Sig.
Szenario 1 (PS)	.08	.58	-	-	-	-	-	-
Szenario 2 (EX)	-	-	-.03	.82	-	-	-	-
Szenario 3 (SK)	-	-	-	-	.01	.93	-	-
Szenario 5 (ES)	-	-	-	-	-	-	.14	.32
Szenario 8 (PS)	.21	.15	-	-	-	-	-	-

Anmerkung. N = 50. Signifikanzniveau nach Bonferroni-Korrektur: $\alpha^{Bonf} = .01$

4.2.5 Prädiktive Validität des SJT

Bevor die Ergebnisse der einzelnen Regressionsanalysen der Single-Item-Skalen aufgeführt werden, soll zuerst auf die Ergebnisse der Voraussetzungsprüfung eingegangen werden. Die Normalverteilung der Residuen war nach visueller Einschätzung der

Histogramme (siehe Anhang A) für die Szenarien 1, 2 und 8 gegeben. Bei den Szenarien 3 und 5 wichen die Werte von einer Normalverteilung ab. Jedoch gilt auch bei der Normalverteilung der Residuen der Zentrale Grenzwertsatz, der für Stichproben mit n > 30 von einer Normalverteilung ausgeht (Urban & Mayerl, 2018). Bei Verletzung der Normalverteilung werden ausserdem die Koeffizienten nicht verfälscht, sondern nur die geschätzten Standardfehler und damit die angegebenen Konfidenzintervalle (Rudolf & Müller, 2004), was bei der Interpretation der Ergebnisse beachtet werden muss. Da die Durbin-Watson-Werte alle im Bereich zwischen 1.5 bis 2.5 lagen, wurde von voneinander unabhängigen Fehlerwerten ausgegangen (Schneider & Ruoff, 2010). Ein Verdacht auf Multikollinearität konnte ausgeschlossen werden, da der Varianz-Inflations-Faktor für alle Regressionsanalysen VIF < 10 (höchster VIF = 1.91) war. Die Varianz der Residuen wurde mittels Streudiagramm der standardisierten Residuen und der standardisierten geschätzten Werten überprüft (siehe Anhang A). Da die Wertebereiche durch die Skalen stark eingeschränkt waren, die Randbereiche eher weniger Werte aufwiesen und somit nicht so stark streuen können, war die Streuung auf den ersten Blick nicht ganz gleichmässig. Es war aber kein Muster erkennbar, deshalb wurde von einer Homoskedastizität ausgegangen (Bortz & Schuster, 2010).

Die Ergebnisse der fünf gerechneten linearen Regressionsanalysen werden in Tabelle 12 zusammengefasst. Fett gedruckt sind jeweils die Werte für das Gesamtmodell. Keiner der fünf Prädiktoren konnte den Wert der jeweils zugewiesene Primärskala des ISK signifikant vorhersagen. Der einzige signifikante Effekt konnte bei Szenario 8 festgestellt werden. Dabei erzielten diejenigen Bewerbenden, die die zweitbeste Antwort im SJT gegeben haben, einen signifikant tieferen Wert (-3.66 Einheiten) in der Prosozialität-Skala im Vergleich mit Bewerbenden, die die passendste Antwort im SJT gegeben haben. Im Gesamtmodell konnten damit 7.9% mehr Varianz erklärt werden als durch Zufall. Jedoch wurde der Effekt im Gesamtmodell knapp nicht signifikant (p = .054). Ausserdem war dieser Effekt nach der Bonferrroni-Korrektur ebenfalls nicht mehr signifikant. Bei Szenario 1, das

laut Fachleute ebenfalls mit Prosozialität in Verbindung steht, wurde ein ähnliches Ergebnis gefunden. Diejenigen Bewerbenden, die die zweitbeste Antwort im SJT gegeben haben, haben einen um -1.82 Einheiten tieferen Wert bei der Prosozialitätsskala als Bewerbende, die die passendste Antwort gegeben haben. Das Gesamtmodell kann mit der untersuchten Stichprobe 4.5% mehr Varianz erklären als der Zufall. Dieser Effekt ist mit p = .062 verhältnismässig nahe am Signifikanzniveau, wird aber ebenfalls bereits vor der Bonferroni-Korrektur nicht signifikant und kann deshalb nicht verallgemeinert werden. Da das korrigierte Bestimmtheitsmass R^2 bei den Prädiktoren 2, 3 und 5 sogar negative Werte annimmt, ist davon auszugehen, dass diese Modelle gar keinen prognostischen Wert haben.

Tabelle 12

Prädiktive Validität der SJT-Items für die entsprechenden ISK-Primärskalen

Prädiktor	Kriteriumsvariable	ß	SE	R^2	korr. R^2	Sig.
Szenario 1	**Prosozialität**		**2.324**	**.084**	**.045**	**.126**
Dummy_zweit		-1.82				.062
Dummy_dritt		0.71				.563
Szenario 2	**Extraversion**		**3.050**	**.008**	**-.057**	**.948**
Dummy_zweit		0.04				.964
Dummy_dritt		1.24				.581
Dummy_viert		0.74				.813
Szenario 3	**Selbstkontrolle**		**2.139**	**.005**	**-.060**	**.974**
Dummy_zweit		-0.25				.766
Dummy_dritt		0.00				1.000
Dummy_viert		-0.38				.720
Szenario 5	**Emotionale Stabilität**		**2.482**	**.020**	**-.044**	**.814**
Dummy_zweit		-0.54				.489
Dummy_dritt		0.59				.699
Dummy_viert		-0.91				.617
Szenario 8	**Prosozialität**		**2.283**	**.116**	**.079**	**.054**
Dummy_zweit		-3.66				.032
Dummy_dritt		-1.23				.194

Anmerkung. N = 50; ß = standardisierter Regressionskoeffizient; SE = Standardfehler des Schätzers; R^2 = quadriertes Bestimmtheitsmass; korr. R^2 = korrigiertes Bestimmtheitsmass. Fett gedruckt sind jeweils die Werte für das Gesamtmodell.

5 Diskussion

Im ersten Teil der Diskussion sollen entlang des angewandten sequenziellen Mixed-Methods-Studiendesigns die Ergebnisse aus Kapitel 4 diskutiert und mithilfe der Theorie begründet und eingeordnet werden. Nach einem Hinweis auf allgemeine Limitationen und Empfehlungen für künftige Forschung, wird mit praktischen Schlussfolgerungen und einem Fazit abgeschlossen.

5.1 Interpretation der qualitativen Ergebnisse

Im qualitativen Teil der Masterarbeit wurde ein Situational Judgment Test erarbeitet, der als Grundlage für das angestrebte Online Self-Assessment dienen sollte. In den folgenden Abschnitten wird die qualitative Entwicklung des SJT diskutiert und damit Leitfrage 1 und 2 beantwortet.

Leitfrage 1
Welche Schlüsselszenarien aus dem Alltag von Verkaufsmitarbeitenden der Gelateria di Berna werden von den Fachleuten als relevant bewertet und wie lauten mögliche Verhaltensalternativen dazu?

Insgesamt wurden 25 Schlüsselszenarien aus dem Alltag von Verkaufsmitarbeitenden der GdB als relevant bewertet, wobei mit 9 davon weitergearbeitet wurde. Die Mehrheit der Szenarien zeigt eine Interaktion mit Gästen. Die zugehörigen Verhaltensalternativen stellen die Befragten vor ein Dilemma, welches in den meisten Fällen zeitliche oder hygienebezogene Aspekte abbildet, die gegen kundenorientiertere Verhaltensweisen abgewogen werden sollen. Eine detaillierte Übersicht über alle finalen Schlüsselszenarien inklusive den dazugehörigen Verhaltensalternativen findet sich in Tabelle 5.

Besonders auffällig war, dass von den ursprünglich 25 gefundenen Arbeitssituationen im ersten qualitativen Workshop 16 Szenarien aufgrund der im Methodenteil definierten Kriterien ausgeschlossen werden mussten. Während alle 25 Arbeitssituationen zwar als relevant eingeschätzt wurden (Leitfrage 1), mussten 8 davon ausgeschlossen werden, weil spezifisches Firmenwissen nötig gewesen wäre, um die passendste Antwort zu finden. Dies

zeigt unter anderem, wie jobspezifisch Arbeitssituationen sein können und weshalb kein SJT dem anderen gleicht. Die acht ausgeschlossenen Arbeitssituationen könnten aber alternativ beispielsweise im Rahmen der On-Boarding-Phase des Personalbeschaffungsprozesses für Schulungen von neuen Mitarbeitenden genutzt werden, um firmenübliche Abläufe zu demonstrieren und zu schulen. Hamilton et al. (2020) konnte für Schulungen mit VR-Umgebungen schon positive Lerneffekte zeigen. Somit könnte nicht nur zusätzlich eine zeitliche und finanzielle Einsparung zu einem anderen Zeitpunkt im Personalbeschaffungsprozess der GdB (Adarve-Gómez et al., 2019), sondern laut (Rogers, 2019) auch mehr Zufriedenheit bei den Mitarbeitenden erreicht werden. Ebenfalls zum Ausschluss von Situationen kam es wegen Fehlens eines Dilemmas für die befragte Person in zwei Szenarien, welches von Bledow und Frese (2009) als nötige Bedingung zur Unterscheidung zwischen Personen mit niedriger und hoher Ausprägung in spezifischen Kompetenzen genannt wird. Bei zwei anderen Szenarien wurde die Fragestellung als zu einfach bewertet, sodass die passendste Verhaltensalternative bezüglich der zugrundeliegenden Kompetenz zu offensichtlich gewesen wäre. Dies ist vor allem beim konstruktbasierten Scoring häufig ein Problem wegen der hohen Transparenz dieses Vorgehens (Hough und Paullin, 1994). Zwar wurde in der Masterarbeit das expertenbasierte Scoring angewendet, jedoch sollten die Fachleute im dritten Workshop die Szenarien aufgrund ausgewählter sozialer Kompetenzen einordnen und umformulieren, was eher einer hybriden Variante, also der Kombination beider Scoring-Typen nahe kommt (Bergman et al., 2006) und deshalb zu leichter durchschaubaren Items führte. Die Kriterien zur Attraktivität und Videotauglichkeit der Arbeitssituationen, welche zum Ausschluss von insgesamt vier weiteren Items führten, wurden besonders hoch gewichtet, da sie für die Zielsetzung der Masterarbeit entscheidend waren. Die empfundene Attraktivität des Stellenprofils und der Unternehmung ist eine der Grundvoraussetzungen dafür, dass sich Stelleninteressierte nach dem Online Self-Assessment für eine Bewerbung entscheiden (Diercks, 2021). Deshalb wurden drei Szenarien ausgeschlossen, die vielleicht für das

Gesamtbild von der Stellenbeschreibung relevant gewesen wären, aber bei denen das Kriterium der Attraktivität stärker gewichtet wurde. Dass die Szenarien videobasiert darstellbar sein mussten, war eine nötige Grundvoraussetzung für das weitere Vorgehen der Masterarbeit.

Leitfrage 2
Welche sozialen Kompetenzen sind besonders wichtig, um in den Schlüsselszenarien adäquat reagieren zu können?

Insgesamt konnten die Fachleute sieben Schlüsselkompetenzen von Verkaufsmitarbeitenden der GdB aus dem Inventar sozialer Kompetenzen (Kanning, 2009) feststellen, die einen Einfluss auf die Arbeitsleistung dieser Stelle haben. Davon konnten Extraversion, Selbstkontrolle und Emotionale Stabilität eindeutig je einem Schlüsselszenario, Prosozialität zwei Schlüsselszenarien zugeordnet werden. So ist laut Fachleute beispielsweise bei Item 1 des entwickelten SJT die Chance höher eine passende Verhaltensweise zu wählen, je ausgeprägter die Prosozialität der befragten Person ist. Der soziale Bezug der fünf Schlüsselszenarien wiederspiegelt die Ergebnisse von Sisson und Adams (2013), die in ihrer Studie sogar 86% der relevanten Kompetenzen im Gastgewerbe im sozialen Bereich ansiedeln. Die übrigen vier gefundenen Schlüsselszenarien wurden trotz fehlendem eindeutigen Bezug zu einer ISK-Skala im SJT beibehalten, da sie ebenfalls relevante Arbeitssituationen aus dem Berufsalltag von Verkaufsmitarbeitenden darstellen. Szenario 4, 6 und 7 widerspiegeln vor allem hygienische Fragestellungen, die in der GdB laut Fachleute einen besonderen Stellenwert haben. Szenario 9 (Feierabendbier der Mitarbeitenden) wurde ausserdem bewusst gewählt, um die im Theorieteil geschilderte Anlockfunktion von Selbstselektions-Tools zu implementieren und die Stelle möglichst attraktiv wirken zu lassen (Diercks, 2021). Jedoch konnte keine testende Komponente daraus abgeleitet werden, weshalb dieses Item für die späteren Analysen nicht berücksichtigt wurde.

Mit der Beantwortung der ersten und zweiten Leitfrage und unter Berücksichtigung des expertenbasierten Scorings der Verhaltensalternativen in Workshop 3 wurde im qualitativen Teil ein kompetenzbasierter, jobspezifischer und textbasierter SJT entwickelt. Grundsätzlich wäre es auch in dieser Form möglich gewesen, den Stelleninteressierten der GdB einen Einblick in den Berufsalltag im Verkauf zu geben. Doch im Hinblick auf die Zielsetzung war eine Übersetzung des textbasierten SJT in ein videobasiertes Online Format sinnvoll. Zum einen wird generell die Realitätsnähe von videobasierten Stimuli in einem SJT höher eingeschätzt (high fidelity, Weekley et al., 2015). So wird in den Aufnahmen für die GdB nicht nur die typische Arbeitsumgebung gezeigt, sondern beispielsweise auch der typische Lärm durch die laute Gelato-Vitrine im Hintergrund wahrgenommen. Dadurch werden verschiedene Sinnessysteme angesprochen, was laut Webster (2016) zu einem realitätsnäheren Erlebnis führt. Die 360°-Verfilmung wurde ausserdem gegenüber einer Verfilmung mit einer normalen Videokamera bevorzugt, da 360°-Videos einen interaktiveren und damit spielerischen Charakter ermöglichen. So konnte mit Item 6 ein Suchbild umgesetzt werden, bei dem die Befragten selbständig die nächste sinnvolle Aufgabe im 360°-Video finden sollten. Ein weiteres Element, das zur Erhöhung des spielerischen Charakters im Zwischenschritt eingebaut wurde, war die Hintergrundgeschichte. Die Anordnung der Videos wurde so gewählt, dass es einen kompletten Einsatz von der Einsatz-Anfrage, über die Begrüssung einer Mitarbeiterin vor Ort, die Bedienung unterschiedlicher Kundinnen und Kunden bis zum anschliessenden Feierabendbier simulierte. Die Kombination aus Interaktivität, anspruchsvollen Aufgaben und einer Storyline kann nicht nur zu grösserem Engagement und einem positiveren Arbeitgeberimage führen (Fetzer et al., 2017; Gkorezis et al., 2021), die präsentierten Inhalte werden dadurch auch intensiver verarbeiten, was wiederum die Selbstselektion unterstützt (Ott et al., 2017). Verschiedene Studien konnten zudem zeigen, dass auch die Akzeptanz gegenüber videobasierten SJT bei den Bewerbenden höher ist als bei textbasierten SJT (z.B. Bardach et al., 2021).

5.2 Interpretation der quantitativen Ergebnisse

Zwar wurde im qualitativen Teil der Masterarbeit bereits ein Instrument entwickelt, jedoch gäbe es ohne anschliessende quantitative Teilstudie weder ein quantifiziertes Referenzprofil (Leitfrage 3), mit welchem die Antworten der Bewerbenden verglichen werden könnten, noch wären Aussagen zur Verständlichkeit oder Validität des Instruments möglich (Leitfrage 4). Hier zeigt sich insbesondere der Nutzen des Mixed-Methods-Design zur Entwicklung eines Instruments verglichen mit monomethodischen Studiendesigns. In den folgenden Abschnitten sollen zunächst die quantitative Stichprobe und die deskriptiven Ergebnisse diskutiert werden, bevor die Leitfragen 3 und 4 beantwortet werden.

5.2.1 Quantitative Stichprobe

Bei der Auswahl der Stichprobe zur Testung des entwickelten SJT wurde dem Forschungsauftrag von Fominykh und Prasolova-Førland (2019) nachgekommen, das entwickelte Instrument im realen Setting zu testen, also im vorliegenden Fall mit realen Bewerbenden der GdB. Dennoch ist die gewonnene Stichprobe nur ein Ausschnitt der Grundgesamtheit aller Bewerbungen der GdB, was bei der Interpretation beachtet werden muss. Ebenfalls könnten die gesetzten Anreize einen Einfluss auf die Teilnahme an der freiwilligen Online Umfrage gehabt haben. Es wurden Gelato-Gutscheine und ein persönliches Feedback zu den Resultaten versprochen. Zum einen könnten Bewerbende sich einfach nur durch die Umfrage durchgeklickt haben für die Gutscheine, zum anderen könnte es sein, dass hauptsächlich besonders ambitionierte Bewerbende die Online Umfrage ausgefüllt haben, welche sehr interessiert an der Stelle waren. Ein Indiz für die letztere Vermutung könnten die hohen Werte im Summenscore des SJT sein. Gegen diese Vermutung sprechen die eher durchschnittlich ausfallenden Werte beim ISK. Was die demografischen Merkmale der Stichprobe betrifft, wird nach Rücksprache mit den Fachleuten davon ausgegangen, dass das Geschlechterverhältnis, Durchschnittsalter und die hauptsächliche Beschäftigung der Bewerbenden nebenher auch der Verteilung der Mitarbeitenden der GdB entsprechen und somit ein typisches Bild der Grundgesamtheit

abgeben. Gruppeneffekte aufgrund der demografischen Merkmale konnten keine gefunden werden, wobei für die t-Tests aufgrund des starken Ungleichgewichts der Verteilungen in den Gruppen nicht alle Voraussetzungen erfüllt werden konnten und diese Einschätzung deshalb mit Vorbehalt zu betrachten ist.

5.2.2 Deskriptive Analyse des SJT

Häufig wird in Studien nicht direkt im realen Setting getestet, sondern es handelt sich oftmals um sogenannte Gelegenheitsstichproben mit gut erreichbaren Personen wie Studierende (Döring & Bortz, 2016). Auch wenn die Online Umfrage der Masterarbeit anonym war, war aufgrund des echten Bewerbungssettings zu erwarten, dass die Bewerbenden sich von ihrer besten Seite zeigen wollten und trotz der gewählten Would-Do-Instruktion (Wie würdest du reagieren?) eher diejenige Verhaltensalternative auswählten, die mit einer höheren sozialen Erwünschtheit in Verbindung gebracht wurde (Wie sollte man reagieren?; Ott et al., 2017). Diese Vermutung wird von der Itemanalyse des SJT gestützt. Die hohen Item-Mittelwerte und die rechtssteile Verteilung aller acht berücksichtigten Items deuten auf hohe psychometrische Item-Schwierigkeiten hin (Bühner, 2011). Viele Bewerbende haben also jeweils die passendste Verhaltensalternative gewählt. Ein weiterer Grund für die hohen Item-Mittelwerte und damit auch die hohen Summenscores im SJT insgesamt könnte die methodischen Vorgehensweise beim Scoring sein. Zum einen muss beachtet werden, dass Items, die leichter übereinstimmend von Fachleuten bewertet werden, auch eher mit sozialer Erwünschtheit zusammenhängen und dadurch verfälschungsanfälliger sind (Krokos et al., 2004, zitiert nach Whetzel & McDaniel, 2009, S. 196). Ployhart und MacKenzie (2010) sowie Muck (2013) raten ausserdem von einem Verhaltenskontinuum (behaviorally uniform), also nur kleinen kontinuierlichen Anpassungen zwischen den Antworten, bei der Gestaltung von Verhaltensalternativen ab, weil diese Vorgehensweise durch die hohe Transparenz ebenfalls sehr verfälschungsanfällig ist. Vielmehr rät die Autorenschaft dazu, ganz verschiedene Verhaltensalternativen anzubieten, die auf den ersten Blick nichts miteinander zu tun haben. Das Verhaltenskontinuum ist im

entwickelten SJT vor allem bei den Items 4 und 7 besonders deutlich ausgeprägt. Die Item-Mittelwerte dieser zwei Items waren im Vergleich mit den meisten anderen Items wiederum verhältnismässig tief, was dieser Theorie wiedersprechen würde. Wäre als Antwortformat ein Rating jeder Verhaltensalternative gewählt worden, wären laut Ployhart und Ehrhart (2003) im Vergleich zum Forced Choice-Format höhere Varianzen erreichbar gewesen. Die Items 1, 4 und 8 schöpfen zudem nicht die komplette Bandbreite der Verhaltensalternativen aus, da jeweils niemand bei diesen Items die unpassendste Antwort gewählt hat. Laut Bühner (2011) ist es problematisch, wenn die Befragten die Bandbreite des Antwortformats nicht vollständig ausnutzen. Dies ist laut Autor ein Hinweis darauf, dass entweder das Antwortformat zu differenziert für die spezifische Zielpopulation ist oder dass das Item psychometrisch zu schwierig oder zu leicht ist. Im vorliegenden Fall wird aufgrund der vorangehenden Argumentation vermutet, dass bei den betreffenden Items mindestens die unpassendste Antwort zu offensichtlich war. Ausserdem hatten schon Rammstedt, Koch, Borg und Reitz (2004) den Verdacht, dass Single-Item-Skalen vergleichsweise anfälliger für sozial erwünschtes Antwortverhalten, was auch zu den vorliegenden Ergebnissen passt. In diesem Zusammenhang soll aber auch darauf hingewiesen werden, dass der SJT im späteren Gebrauch als Selbstselektionstool angewendet werden soll und die Ergebnisse dabei nur von den Stelleninteressierten selbst gesehen werden. So wird davon ausgegangen, dass die Summenscores des SJT tiefer ausfallen werden, wenn das Online Self-Assessment nur zu persönlichen Zwecken genutzt wird (Ott et al., 2017).

5.2.3 Deskriptive Analyse des ISK

Die Mittelwerte der ISK-Skalen der Stichprobe lagen verglichen mit der Normierungsstichprobe von Kanning (2009) im Durchschnitt. Jedoch waren die internen Konsistenzen zum Teil etwas tiefer als in der Normierungsstichprobe und teilweise im sehr niedrigen Bereich (Primärskala Handlungsflexibilität), was für eine tiefe Messgenauigkeit bei der vorliegenden Stichprobe spricht (Janssen & Laatz, 2017). Da die ISK-Ergebnisse laut Kanning (2009) erst aussagekräftig sind, wenn sie in Relation zu einem gewünschten

Referenzprofil betrachtet werden, wurden die Antworten von «idealen» Verkaufsmitarbeitenden von den Fachleuten der GdB simuliert und damit auch Leitfrage 3 beantwortet.

Leitfrage 3
Wie würden «ideale» Verkaufsmitarbeitende in den Schlüsselszenarien reagieren?

Es ist auffällig, dass laut Einschätzung der Fachleute ideale Verkaufsmitarbeitende in allen Primärskalen bis auf die Durchsetzungsfähigkeit, Konfliktbereitschaft und Kompromissbereitschaft Standardwerte in überdurchschnittlicher Höhe (> 110) mitbringen sollten. Dieses Ergebnis bestätigt nochmals, dass die sozialen Kompetenzen im Gastgewerbe einen hohen Stellenwert haben (Sisson & Adams, 2013). Bei den drei im Referenzprofil verhältnismässig schwach ausgeprägten sozialen Kompetenzen Durchsetzungsfähigkeit, Konfliktbereitschaft und Kompromissbereitschaft ist zu vermuten, dass diese in schwacher Ausprägung im Gastgewerbe effektiver sind zur Aufgabenbewältigung. Die Selbsteinschätzung der Bewerbenden lag grösstenteils im durchschnittlichen Bereich zwischen 100 bis 110, wies aber eine sehr ähnliche Form wie das Referenzprofil auf. Lediglich die Ausprägungen des Referenzprofils schlugen stärker in Richtung der Minimal- und Maximalwerte aus. Die ähnliche Form der Profile könnte damit erklärt werden, dass ein Grossteil der Befragten eine Vorstellung davon hatte, welche Antworten für die betreffende Stelle erwünscht wären (Should-Do) und dementsprechend geantwortet hat. Dass die Werte der Bewerbenden aber verglichen mit der Normierungsstichprobe nicht überdurchschnittlich hoch ausfielen, stützt nochmals die Theorie von Rammstedt et al. (2004), dass Multi-Item-Skalen weniger anfällig für sozial erwünschtes Verhalten sind.

5.2.4 Validität des SJTs

Allgemein wurde bei den Korrelations- und Regressionsanalysen zur Einschätzung der Validität des SJTs streng vorgegangen. Es wurden nicht-parametrische Tests

durchgeführt und mit Bonferroni-Korrekturen gerechnet. Deshalb fallen berechnete Zusammenhänge mit höherer Wahrscheinlichkeit nicht signifikant aus. Ein hoher α-Fehler, also dass Zusammenhänge fälschlicherweise signifikant werden (Armstrong, 2014), sollte auf Kosten eines höheren β-Fehlers (dass signifikante Zusammenhänge unentdeckt bleiben) aber vermieden werden. Grundsätzlich war die Ausgangslage für die quantitativen Analysen nicht ganz optimal, da im Scoring der Fachleute nur ein bis zwei Items pro Primärskala zugeteilt wurden und mit insgesamt neun Items im SJT nicht viel Raum für den Ausschluss von unpassenden Items blieb. Daraus entstanden unterschiedliche Herausforderung, die es methodisch aufzufangen galt, aber sich dementsprechend auf die Resultate auswirkten.

Als Voraussetzungsprüfung für die linearen Regressionen zur Beantwortung von Leitfrage 4 wurden die Zusammenhänge zwischen den fünf kompetenzbasierten Items mit den jeweiligen Primärskalen berechnet. Insgesamt wurden nur sehr kleine Zusammenhänge zwischen den beiden Instrumenten festgestellt. Der stärkste Zusammenhang wurde für Item 8 («Das Gelato fällt auf den Boden») mit der Prosozialität-Skala gefunden, wobei es sich gemäss Kuckartz et al. (2013) um einen kleinen bis mittleren Effekt handelte. Dieser war aber wie alle anderen Zusammenhänge als zufälliger, stichprobenabhängiger Effekt zu bewerten, da die Zusammenhänge schon vor der Bonferroni-Korrektur nicht signifikant wurden. Mit dieser Ausgangslage sollte in den nachfolgenden linearen Regressionen Leitfrage 4 beantwortet werden.

Leitfrage 4
Wie ist die prädiktive Validität des 360°-videobasierten Situational Judgment Tests bezogen auf ausgewählte Kompetenzen zu bewerten?

Keines der fünf getesteten Szenarien konnte signifikant zur Aufklärung der Varianz in den entsprechenden ISK-Werten der Bewerbenden beitragen. Die beste Vorhersage konnte mit Szenario 8 erreicht werden. Dabei konnten 7.9% mehr Varianz in den ISK-Werten der Primärskala Prosozialität erklärt werden als durch Zufall. Doch der Effekt wurde schon vor der Bonferroni-Korrektur nicht signifikant und ist somit nicht verallgemeinerbar. Szenario 1,

welches ebenfalls Werte der Prosozialität vorhersagen sollte, konnte in der vorliegenden Stichprobe lediglich 4.5% der Varianz erklären. Für Szenario 2, 3 und 5 konnte kein prognostischer Wert gefunden werden.

Ein möglicher Grund für die schwachen Zusammenhänge könnte sein, dass laut Kanning (2009) die soziale Kompetenz ein multidimensionales Konstrukt ist, das sich aus sozial kompetentem Verhalten zusammensetzt. So misst der entwickelte SJT streng genommen nicht soziale Kompetenzen direkt, sondern ist nur eine Einschätzung dieser aufgrund des beobachteten Verhaltens. Ein weiterer Grund für die ungenügende prädiktive Validität könnte sein, dass andere Studien, die eine gute prädiktive Validität für SJTs zeigen konnten, mehrheitlich die Job Performance als Kriteriumsvariable in die Analyse aufgenommen, die mittels Vorgesetztenbeurteilungen gemessen wurde (Lievens & Sackett, 2012; Webster et al., 2020). Dadurch wies die Kriteriumsvariable in diesen Studien relativ konkrete, jobspezifische Merkmale auf, die auch in den für die Stelle entwickelten SJTs simuliert wurden. Die Ähnlichkeit der Prädiktor- und Kriteriumsvariable könnte dabei zur gefundenen prädiktiven Validität beigetragen haben. Da in der vorliegenden Masterarbeit die ISK-Werte vorhergesagt werden sollten, wurden mit einem sehr jobspezifischen SJT die Werte eines sehr allgemein formulierten Instruments prognostiziert.

Kanning (2009) weist zudem darauf hin, dass die soziale Kompetenz als multidimensionales Konstrukt auch ein multidimensionales Verfahren zu dessen Messung voraussetzt. Während im ISK pro Primärskala fünf bis neun Items für eine aussagekräftige Einschätzung führen sollen, steht der entwickelte SJT mit ein bis maximal zwei Items pro Primärskala in einem klaren Kontrast da. Laut Diamantopoulos et al. (2012) können Single-Item-Skalen nur unter sehr spezifischen Bedingungen eine vergleichbare oder bessere prädiktive Validität erreichen als Multi-Item-Skalen. Als Daumenregel geben sie an, dass Single-Item-Skalen vor allem bei kleinen Stichproben (< 50), kleinen erwarteten Effektgrössen (< .30), hohen Inter-Item-Korrelationen (< .80) und semantischer Redundanz

als Alternative zu Multi-Item-Skalen in Erwägung gezogen werden können. Mit den tiefen Inter-Item-Korrelationen fällt der entwickelte SJT dabei deutlich aus diesem Raster.

Ein weiterer möglicher Grund dafür, dass weniger gute Prognosefähigkeit des entwickelten SJT könnte die einseitige Betrachtung der Tätigkeiten von Verkaufsmitarbeitenden sein. In der Metaanalyse von McDaniel et al. (2007) konnten SJT basierend auf einer Tätigkeitsanalyse im Vorfeld höhere prädiktive Validitäten erreichen als SJT ohne Tätigkeitsanalyse. Möglicherweise hätte eine Tätigkeitsanalyse mit einer objektiven Sicht auf die Aufgaben zu besseren prädiktiven Werten geführt. Jedoch wären dann möglicherweise seltene, aber relevante Ereignisse, die mithilfe der Critical Incident Technique aufgedeckt werden konnten, bei der Tätigkeitsanalyse eines «normalen» Arbeitstags unentdeckt geblieben (Christian et al., 2010).

5.3 Limitationen und zukünftige Forschung

Neben den bereits genannten möglichen Einflüssen der methodischen Vorgehensweisen bei der Interpretation der Ergebnisse soll zum Schluss noch einige allgemeine Limitationen der vorliegenden Masterarbeit hingewiesen werden. Was im Nachhinein als eine Limitation angesehen wird, sind die relativ streng gesetzten Ausschlusskriterien in Workshop 1. Diese hatten einen bedeutenden Einfluss auf die späteren quantitative Analysen, da durch die Reduktion von insgesamt 25 Items auf schlussendlich 9 Items nur 1-2 Items pro Schlüsselkompetenz gefunden werden konnten. Mit mehreren Items pro Primärskala wären andere quantitative Analysen möglich gewesen. Die linearen Regressionen hätten nicht unbedingt mit Dummyvariablen als Prädiktor gerechnet werden müssen, da der Summenscore einer Multi-Item-Skala unter bestimmten Voraussetzungen als Intervallskala behandelt werden darf (Bortz & Schuster, 2010). Für weitere Forschungsbestreben in diesem Bereich wird deshalb empfohlen, die Ausschlusskriterien nicht zu streng zu setzen und mit nachfolgenden quantitativen Kriterien wie beispielsweise einer Faktoranalyse zwischen passenden und weniger passenden Items

zu unterscheiden. Alternativ hätten die Fachleute auch zuerst die Schlüsselkompetenzen ausarbeiten können und aufgrund dieser solange passende Arbeitssituationen erarbeiten können, bis es für alle Schlüsselkompetenzen genügend Items gehabt hätte. Jedoch wäre dieses Vorgehen weniger explorativ gewesen und sicherlich wäre dann das eine oder andere relevante Szenario nicht genannt worden, da es nicht zu den Schlüsselkompetenzen gepasst hätte.

Eine weitere Limitation der vorliegenden Arbeit ist, dass keine Aussage zur internen Konsistenz möglich ist, da es sich um eine Querschnittsstudie handelt. Zur Einschätzung der internen Konsistenz von Single-Item-Skalen empfehlen Beierlein et al. (2014) die Retest-Methode, jedoch wäre dazu ein zweiter Messzeitpunkt nötig gewesen, was im vorliegenden Fall wegen des Bewerbungssettings nicht sinnvoll gewesen wäre und auch aus zeitlichen Gründen nicht möglich war. Ebenfalls als Limitation zu nennen ist, dass der entwickelte SJT direkt verfilmt wurde und über die Validität des textbasierten SJT keine Aussagen möglich sind. Georgiou et al. (2019) haben in ihrer Studie zuerst die Konstrukt- und Kriteriumsvalidität des textbasierten SJT überprüft, bevor sie eine gamifizierte Version daraus erstellt haben. Damit wären gerade im Hinblick auf das Forschungsdefizit zu Validitäten von Online Formaten aufschlussreichere Aussagen zu den Unterschieden der verschiedenen Formate möglich gewesen. Dies hätte jedoch den Rahmen der vorliegenden Masterarbeit gesprengt.

Zum Schluss soll noch auf eine technische Limitation hingewiesen werden. In der durchgeführten Online Umfrage wurden auf der Seite mit dem Link zur Weiterleitung zum SJT am meisten Abbrüche (75%) verzeichnet. Ein möglicher Grund dafür könnte sein, dass Bewerbende sich die Videos angeschaut und sich dann gegen eine Bewerbung entschieden haben, ohne die Befragung abzuschliessen. Jedoch ist dies nur eine Annahme Es könnten auch technische Probleme bei der Weiterleitung zum SJT aufgetreten sein. Gansser und Zimmermann (2017) weisen darauf hin, dass die Abbruchrate insbesondere bei Nichtfunktionalität auf Smartphones massiv erhöht wird. Ausserdem kann es sein, dass

die Befragten nach Abschluss des SJT den Weg nicht mehr zurück in die Umfrage gefunden haben. Deshalb sollte bei einer zukünftigen Umfrage ähnlicher Art darauf geachtet werden, dass alle Teile der Umfrage im gleichen Umfrage-Tool integriert werden können.

Um weitere Erkenntnisse zur Validität von Online Self-Assessments im Personalbeschaffungsprozess zu erhalten, sollte zukünftige Forschung Überlegungen anstreben, wie mit dem jobspezifischen Charakter solcher Formate umgegangen werden soll. Viele relevante Arbeitssituationen der GdB konnten nicht in das Online Self-Assessment einfliessen, weil spezifisches Firmenwissen nötig gewesen wäre, um die gewünschte Verhaltensalternative zu finden. Es wird deshalb die Vermutung aufgestellt, dass je spezifischer das Instrument auf eine Stelle oder eine Unternehmung ausgerichtet wird, desto weniger Personen können die effektivsten Verhaltensalternativen noch finden. Diese Vermutung wird gestützt von einer Studie von Motowidlo und Beier (2010). Dabei konnte gezeigt werden, dass Personen ohne Vorerfahrungen in den relevanten Anforderungen nur in ihrem generellen Wissen getestet werden. Personen, die bereits ähnliche Berufserfahrung mitbringen, werden bei entsprechenden Fragestellungen auch in ihrem spezifischen, berufsrelevanten Wissen getestet und deshalb sind dabei bessere prädiktive Ergebnisse zu erwarten. Ein Online Self-Assessment, welches in der Attraction-Phase des Personalbeschaffungsprozess zum Einsatz kommt, sollte sicher nicht zu schwierig gestaltet sein, um nicht zu viele qualifizierte Stelleninteressierte abzuschrecken. Wenn es hingegen zu einfach oder allgemein gestaltet wird, so verliert es seinen prädiktiven Charakter.

5.4 Praktische Schlussfolgerungen

Aufgrund der Ergebnisse der vorliegenden Masterarbeit lassen sich einige Schlussfolgerungen für die Praxis ableiten. Verhoeven (2020) betont die Wichtigkeit, dass sich Arbeitgebende ein ausreichendes Verständnis darüber aneignen müssen, was sich

Bewerbende von Unternehmen wünschen. Zwar wurden die Bewerbenden der GdB nicht gefragt, was sich im Bewerbungsprozess wünschen. Fominykh & Prasolova-Førland (2019) haben in ihrer Befragung von jungen Stellensuchenden aber feststellen können, dass diese vor allem eine wirkliche Veranschaulichung von Arbeitsplätzen, Arbeitsaufgaben sowie den dazu erforderlichen Kompetenzen mit einem Feedback zu ihrer Leistung fordern. Mit der Entwicklung des Online Self-Assessments wurde ein Instrument geschaffen, das mit neuster Technologie wie 360°-Videos und Game Elementen sowie mithilfe etablierter Verfahren der Personalauswahl (ISK, SJT) einiger dieser Forderungen nachkommen ist. Denn es wird damit der Zugang zu Informationen erleichtert, die bisher nur über Arbeitsproben sowie Erläuterungen in einem Interview oder einer Präsentation erreichbar waren. Ausserdem bestätigen die guten Bewertungen und die Bemerkungen der Befragten in der Online Umfrage, dass das Online Self-Assessment ein Schritt in die richtige Richtung war: «Die Videos waren richtig gut! Sehr lebensnah! Und ich musste ein paar Mal laut lachen - es hat auch Spass gemacht» oder «Ich habe die Videos als Beispiele sehr authentisch gefunden und konnte mich auch aufgrund der 360°-Ansicht gut in die Situation begeben. Eine interessante Art von Umfrage!». Wie der Bericht des Talent Board (2021) zeigte, ist die Chance gross, dass positive Erfahrungen im Rahmen der Candidate Experience mit dem Umfeld und der Öffentlichkeit geteilt werden. Damit werden die Erfolgschancen der GdB auf dem Arbeitsmarkt durch Weiterempfehlung und Aufmerksamkeit erhöht. Dieses Ergebnis soll andere Unternehmen des Gastgewerbes ermutigen, ähnliche Formate auszuprobieren und ihre Erfahrungen zu teilen.

Dennoch dürfen die «Gefahren» des spielerischen Charakters des Instruments nicht ausser Acht gelassen werden. Fominyykh und Prasolova-Førland (2019) warnen davor, dass gerade bei Simulationen dieser Art die Gefahr besteht, dass ein unrealistischer Spass-Faktor eines Jobprofils entsteht, der wiederum zu falschen Karriereentscheidungen und damit verbundenem Mehraufwand für Unternehmen führen kann. Deshalb ist es wichtig, dass das Online Self-Assessment nicht als alleinstehendes Instrument genutzt wird,

sondern in Kombination mit anderen Verfahren, wie beispielsweise einem anschliessenden strukturierten Interview zur Anwendung kommt, um die Motivation hinter der Bewerbung abschliessend verstehen zu können (Verhoeven, 2020).

Damit die Stelleninteressierten nach der Durchführung des Online Self-Assessments die Stelle nicht nur erleben konnten, sondern auch eine Einschätzung ihrer Antworten erhalten, muss in einem nächsten Schritt eine Feedback-Funktion eingebaut werden, die aufgrund der Antworten eine entsprechende Rückmeldung entweder zur Leistung oder zu den damit verbundenen Kompetenzen abgibt. Dabei ist zu berücksichtigen, dass die empfundene Fairness und damit auch die eingeschätzte Attraktivität eines Unternehmens von der Rückmeldung des Online Self-Assessments abhängen und dieses deshalb möglichst anforderungsbezogen formuliert werden muss (Anseel & Lievens, 2009). So wird sichergestellt, dass das Feedback den Selbstselektionsprozess hinreichend unterstützt. In den Rückmeldungen der Befragten der quantitativen Befragung wurde ausserdem explizit nach einem Zwischenfeedback nach den Szenarien gefragt, sodass direkt abgeschätzt werden kann, ob die gewählte Verhaltensweise dem gewünschten Profil entspricht oder nicht. Dies könnte als weiteres Element verwendet werden (Fortschrittsüberwachung), um den spielerischen Charakter des Online Self-Assessments weiter auszubauen (Glover, 2013).

5.5 Fazit

Im Rahmen der vorliegenden Masterarbeit wurde ein Online Self-Assessment basierend auf einem 360°-videobasierten Situational Judgment Test für die Gelateria di Berna entwickelt. Unter Berücksichtigung der schwierigen Situation auf dem Arbeitsmarkt und den besonderen Herausforderungen im Gastgewerbe sollte das Instrument nicht nur einen realistischen Einblick in den Berufsalltag von Verkaufsmitarbeitenden geben, sondern durch spielerische Elemente die Neugierde der Stelleninteressierten wecken und dadurch die Candidate Experience positiv beeinflussen. Die positiven Rückmeldungen nach einer

ersten Testung des Instruments im realen Bewerbungssetting deuten darauf hin, dass dieser Zielsetzung nachgekommen wurde. Um die Selbstselektion der Stelleninteressierten zusätzlich zu unterstützen, wurde auch eine testende Funktion in das Instrument eingebaut werden, welche eine Einschätzung der sozialen Kompetenzen der Befragten erlaubt. Es konnte ein Referenzprofil mit den angestrebten Ausprägungen der sozialen Kompetenzen von idealen Verkaufsmitarbeitenden definiert werden. Es konnte aber kein bedeutender Vorhersagewert des entwickelten Online Self-Assessment bezogen auf die sozialen Kompetenzen erreicht werden. Möchte die GdB in Zukunft die Stelleninteressierten aufgrund ihrer sozialen Kompetenzen mit automatischem Feedback in ihrer Selbstselektion unterstützen, bedarf es einer gezielten Überarbeitung und Testung der SJT-Items, bevor die Inhalte verfilmt werden. Es bleibt ausserdem die Frage offen, welche Auswirkungen der Einsatz des entwickelten Online Self-Assessment künftig auf die Zusammensetzung des Bewerbenden-Pools und die anschliessenden Interviews haben wird.

Literaturverzeichnis

Anseel, F. & Lievens, F. (2009). The mediating role of feedback acceptance in the relationship between feedback and attitudinal and performance outcomes. *International Journal of Selection and Assessment*, *17*(4), 362-376. https://doi.org/10.1111/j.1468-2389.2009.00479.x

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*(5), 502-508. https://doi.org/10.1111/opo.12131

Armstrong, M. B., Ferrell, J. Z., Collmus, A. B. & Landers, R. N. (2016). Correcting misconceptions about gamification of assessment: More than SJTs and badges. *Industrial and Organizational Psychology*, *9*(3), 671-677. https://doi.org/10.1017/iop.2016.69

Bardach, L., Rushby, J. V., Kim, L. E. & Klassen, R. M. (2021). Using video-and text-based situational judgement tests for teacher selection: A quasi-experiment exploring the relations between test format, subgroup differences, and applicant reactions. *European Journal of Work and Organizational Psychology, 30*(2), 251-264. https://doi.org/10.1080/1359432X.2020.1736619

Bauer, T. N., Truxillo, D. M., Tucker, J. S., Weathers, V., Bertolino, M., Erdogan, B. & Campion, M. A. (2006). Selection in the information age: The impact of privacy concerns and computer experience on applicant reactions. *Journal of Management, 32*(5), 601-621. https://doi.org/10.1177/0149206306289829

Baur, N. (2011). Das Ordinalskalenproblem. In L. Akremi, N. Baur & S. Fromm (Hrsg.), *Datenanalyse mit SPSS für Fortgeschrittene 1* (S. 211-221). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-93041-1_10

Baur, N., Kelle, U. & Kuckartz, U. (2017). Mixed Methods - Stand der Debatte und aktuelle Problemlagen. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *69*(2), 1-37. https://doi.org/10.1007/s11577-017-0450-5

Beierlein, C., Kovaleva, A., László, Z., Kemper, C. J. & Rammstedt, B. (2014). *Eine Single-Item-Skala zur Erfassung der Allgemeinen Lebenszufriedenheit: Die Kurzskala Lebenszufriedenheit-1 (L-1)*. Köln: GESIS.

Bergkvist, L. & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of marketing research*, *44*(2), 175-184. https://doi.org/10.1509/jmkr.44.2.175

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*(3), 223-235. Oxford: Blackwell Publishing Ltd. https://doi.org/10.1111/j.1468-2389.2006.00345.x

Blanz, M. (2021). *Forschungsmethoden und Statistik für die Soziale Arbeit: Grundlagen und Anwendungen*. Stuttgart: Kohlhammer Verlag.

Bledow, R. & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology, 62*(2), 229-258. https://doi.org/10.1111/j.1744-6570.2009.01137.x

Blickle, G. (2019). Leistungsbeurteilung. In F. W. Nerdinger, G. Blickle & N. Schaper (Hrsg.), *Arbeits- und Organisationspsychologie* (303-323). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-56666-4_18

Black, J. S. & van Esch, P. (2020). AI-enabled recruiting: What is it and how should a manager use it?. *Business Horizons, 63*(2), 215-226. https://doi.org/10.1016/j.bushor.2019.12.001

Bortz, J. & Schuster, C. (2010). Einfache lineare Regression. In C. Schuster (Hrsg.), *Statistik für Human- und Sozialwissenschaftler* (183-202). Berlin, Heidelberg: Springer Verlag.

Brenner, F. (2016). Bridging the Scientist-Practitioner Gap: Einflussfaktoren auf die Bewerberakzeptanz bei neuen Technologien am Beispiel zeitversetzter Video-Interviews. In T. Verhoeven (Hrsg.), *Candidate Experience* (S. 71-89). Wiesbaden: Springer Gabler.

Brown, M. I. & Grossenbacher, M. A. (2017). Can you test me now? Equivalence of GMA tests on mobile and non-mobile devices. *International Journal of Selection and Assessment*, *25*(1), 61-71. https://doi.org/10.1111/ijsa.12160

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion.* Hallbergmoos: Pearson.

Bundesamt für Statistik BFS (2022, 30. Mai). *Beschäftigungsbarometer im 1. Quartal 2022* [Medienmitteilung]. Verfügbar unter: https://www.bfs.admin.ch/bfs/de/home/aktuell/medienmitteilungen.assetdetail.22604245.html

Chapman, D. S. & Mayers, D. (2015). Recruitment processes and organizational attraction. In I. Nikolaou & J. K. Oostrom (Hrsg.), *Employee recruitment, selection and assessment. Contemporary issues for theory and practice* (S. 27-42). London: Psychology Press. https://doi.org/10.4324/9781315742175

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*(1), 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Cook, R., Jones-Chick, R., Roulin, N. & O'Rourke, K. (2020). Job seekers' attitudes toward cybervetting: Scale development, validation, and platform comparison. *International Journal of Selection and Assessment, 28*(4), 383-398. https://doi.org/10.1111/ijsa.12300

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P. & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*, *40*(3), 434-449. https://doi.org/10.1007/s11747-011-0300-3

Diercks, J. & Kupka, K. (2013). *Recrutainment: Spielerische Ansätze in Personalmarketing und -auswahl*. Wiesbaden: Springer Gabler. https://doi.org/10.1007/978-3-658-01570-1

Diercks, J. (2021). Online-Assessment. In M. Rütten & K. Bierer (Hrsg.), *Future Talents* (S. 133-153). Wiesbaden: Springer Gabler. https://doi.org/10.1007/978-3-658-33023-1_10

DIN (2016). DIN 33430: Anforderungen an berufsbezogene Eignungsdiagnostik. Berlin: Beuth.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Wiesbaden: Springerverlag.

Fellner, K. (2019). *Moderne Personalauswahl*. Wiesbaden: Springer Fachmedien.

Fetzer, M., McNamara, J. & Geimer, J. L. (2017). Gamification, serious games and personnel selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore & C. Semedo (Hrsg.), *The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention* (S. 293-309). Hoboken, NY: Wiley Blackwell.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4), 327-358.

Fominykh, M. & Prasolova-Førland, E. (2019). Immersive job taste: A concept of demonstrating workplaces with virtual reality. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (S. 1600-1605). IEEE. DOI: 10.1109/VR.2019.8798356

Gansser, O. & Zimmermann, S. (2017). Online-versus mobile Umfragen in der Marktforschung. In O. Gansser & B. Krol (Hrsg.), *Moderne Methoden der Marktforschung* (S. 73-91). Wiesbaden: Springer Gabler.

Garman, A. N., Johnson, M. P. & Howard, D. M. (2006). *Development of a situational judgment test to assess educational outcomes.* Poster presented at the annual meeting of the Society of Industrial / Organizational Psychologists, Dallas, TX.

Georgiou, K., Gouras, A. & Nikolaou, I. (2019). Gamification in employee selection: The development of a gamified assessment. *International journal of selection and assessment*, *27*(2), 91-103. https://doi.org/10.1111/ijsa.12240

Gkorezis, P., Georgiou, K., Nikolaou, I. & Kyriazati, A. (2021). Gamified or traditional situational judgement test? A moderated mediation model of recommendation intentions via organizational attractiveness. *European Journal of Work and Organizational Psychology*, *30*(2), 240-250. https://doi.org/10.1080/1359432X.2020.1746827

Glover, I. (2013). Play As You Learn: Gamification as a Technique for Motivating Learners. In J. Herrington, A. Couros & V. Irvine (Hrsg.), *Proceedings of Edmedia + Innovate Learning* (S. 1999-2008). Waynesville, NC: Association for the Advancement of Computing in Education (AACE).

Gómez, C. A., Carvajal, D. A. C., Zapata, E. J. R. & Villar-Vega, H. (2019). A review of virtual reality videogames for job-training applications. *Revista CINTEX*, *24*(1), 64-70. https://doi.org/10.33131/24222208.346

Greene, J. C., Caracelli, V. J. & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational evaluation and policy analysis*, *11*(3), 255-274. https://doi.org/10.3102/01623737011003255

Hamilton, D., McKechnie, J., Edgerton, E. & Wilson, C. (2020). Immersive virtual reality as a pedagogical tool in education: A systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education, 8*(1), 1–32. https://doi.org/10.1007/s40692-020-00169-2

Hedderich, J. & Sachs, L. (2020). Statistische Modellbildung. In J. Hedderich & L. Sachs (Hrsg.), *Angewandte Statistik* (S. 815-935). Berlin, Heidelberg: Springer Spektrum. https://doi.org/10.1007/978-3-662-62294-0_8

Hiltmann, M. (2013). Online-Self-Assessments: Ein Impuls zur persönlichen und beruflichen Weiterentwicklung. *Wirtschaftspsychologie*, 1, 72-80.

Höft, S. & Goerke, P. (2014). Traditionelle Arbeits- und Anforderungsanalyse trifft modernen Kompetenzmanagementansatz: Rosenkrieg oder Traumhochzeit. *Wirtschaftspsychologie, 16*(1), 5-14.

Hough, L. & Paullin, C. (1994). Construct-oriented scale construction: The rational approach. In G. S. Stokes, M. D. Mumford & W. A. Owens (Hrsg.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (S. 109–145). Palo Alto, CA: Consulting Psychologists Press, Inc.

IBM Corp. Released 2021. *IBM SPSS Statistics for Windows*, Version 28.0. Armonk, NY: IBM Corp.

Ireland, T. (2016). How companies like uber & siemens are gamifying recruitment. *Taledo Blog.* Verfügbar unter: https://www.taledo.com/blog/uber-siemens-gamification-recruitment

Ito, T. A. & Urland, G. R. (2003). Race and gender on the brain: Electrocortical measures of attention to the race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology, 85*(4), 616–626. https://doi.org/10.1037/0022-3514.85.4.616

Jansen, A., Melchers, K. G. & Kleinmann, M. (2012). Der Beitrag sozialer Kompetenz zur Vorhersage beruflicher Leistung. *Zeitschrift für Arbeits-und Organisationspsychologie, 56*(2), 87-97. Göttingen: Hogrefe Verlag.

Janssen, J. & Laatz, W. (2017). Nicht parametrische Tests. In J. Janssen & W. Laatz (Hrsg.), *Statistische Datenanalyse mit SPSS* (631-692). Berlin: Springer Gabler.

Kalafatoğlu, Y. (2020). Gamification in business: A review of the studies. In M. Bilgin, H. Danis & E. Demir (Hrsg.), *Eurasian Business Perspectives*. *Eurasian Studies in Business and Economics* (S. 53-75). Cham: Springer. https://doi.org/10.1007/978-3-030-52294-0_4

Kanning, U. P. (2009). *ISK - Inventar sozialer Kompetenzen*. Göttingen: Hogrefe.

Kanning, U. P. & Schuler, H. (2014). Simulationsorientierte Verfahren der Personalauswahl. In H. Schuler & U. P. Kanning (Hrsg.), *Lehrbuch der Personalpsychologie* (215-256). Göttingen: Hogrefe Verlag.

Karim, M. N., Kaminsky, S. E. & Behrend, T. S. (2014). Cheating, reactions, and performance in remotely proctored testing: An exploratory experimental study. *Journal of Business and Psychology*, *29*(4), 555-572. https://doi.org/10.1007/s10869-014-9343-z

Katzlinger, E. (2017). Gamification elements and online games in the recruiting process. In M. Pivec & J. Gründler (Hrsg.), *Proceedings of the 11th European Conference on Games Based Learning* (S. 311–319). Reading, UK: Academic Conferences and Publishing International Limited.

King, D. D., Ryan, A. M., Kantrowitz, T., Grelle, D. & Dainis, A. (2015). Mobile internet testing: An analysis of equivalence, individual differences, and reactions. *International Journal of Selection and Assessment*, *23*(4), 382-394. https://doi.org/10.1111/ijsa.12122

Klassen, R. M., Kim, L. E., Rushby, J. V. & Bardach, L. (2019). Can we improve how we screen applicants for initial teacher education?. *Teaching and Teacher Education*, *87*, 102949. https://doi.org/10.1016/j.tate.2019.102949

Koch, D. P. A. (2010). Die Task-Analysis-Tools (TAToo)-Entwicklung, empirische und praktische Prüfungen eines Instrumentes für Anforderungsanalysen (Dissertationsschrift, Technische Universität Dresden, 2010). *Dissertation Abstracts* International. (OCLC-Nummer 725162017).

Konradt, U. & Sarges, W. (2003). *E-Recruitment und E-Assessment: Rekrutierung, Auswahl und Beurteilung von Personal im Inter-und Intranet*. Göttingen: Hogrefe Verlag.

Konradt, U., Warszta, T. & Ellwart, T. (2013). Fairness perceptions in web-based selection: Impact on applicants' pursuit intentions, recommendation intentions, and intentions to reapply. *International Journal of Selection and Assessment, 21*(2), 155-169. https://doi.org/10.1111/ijsa.12026

Kuckartz, U. (2014). *Mixed Methods: Methodologie, Forschungsdesigns und Analyseverfahren*. Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-531-93267-5

Kuckartz, U. (2017). Datenanalyse in der mixed-methods-Forschung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *69*(2), 157-183. https://doi.org/10.1007/s11577-017-0456-z

Kuckartz, U., Rädiker, S., Ebert, T. & Schehl, J. (2013). *Statistik: eine verständliche Einführung*. Wiesbaden: Springer VS.

Kühl, S., Strodtholz, P. & Taffertshofer, A. (2009). *Handbuch Methoden der Organisationsforschung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Kuhlmei, E. (2020). Prüfung der Normalverteilungsannahme von stetigen,(mindestens) intervallskalierten Variablen. In E. Kuhlmei (Hrsg.), *Lerne mit uns komplexe Statistik!* (S.191-206). Berlin, Heidelberg: Springer.

Ladkin, A. & Buhalis, D. (2016). Online and social media recruitment: Hospitality employer and prospective employee considerations. *International journal of contemporary hospitality management, 28*(2), 327-345. https://doi.org/10.1108/IJCHM-05-2014-0218

Landers, R. N. & Sackett, P. R. (2012). Offsetting performance losses due to cheating in unproctored internet-based testing by increasing the applicant pool. *International Journal of Selection and Assessment*, *20*(2), 220-228. https://doi.org/10.1111/j.1468-2389.2012.00594.x

Langer, M., König, C. J. & Krause, K. (2017). Examining digital interviews for personnel selection: Applicant reactions and interviewer ratings. *International journal of selection and assessment*, *25*(4), 371-382. https://doi.org/10.1111/ijsa.12191

Larson, K. (2019). Serious games and gamification in the corporate training environment: A literature review. *TechTrends, 64*, 319-328. https://doi.org/10.1007/s11528-019-00446-7

Le Corff, Y., Gingras, V. & Busque-Carrier, M. (2017). Equivalence of unproctored internet testing and proctored paper-and-pencil testing of the Big Five. *International Journal of Selection and Assessment, 25*(2), 154-160. https://doi.org/10.1111/ijsa.12168

Lee, E. A. L. & Wong, K. W. (2008). A review of using virtual reality for learning. In Z. Pan, A. D. Cheok, W. Müller & A. El Rhalibi (Hrsg.) *Transactions on edutainment I* (231-241). Berlin, Heidelberg: Springer.

Leeds, J. P., Griffith, R. & Frei, R. L. (2003). The Development and Validation of a Situational Judgement Test for Security Officers. *Security Journal, 16*(1), 63-78. https://doi.org/10.1057/palgrave.sj.8340126

Lievens, F. (2000). Development of an empirical scoring scheme for situational inventories. *European Review of Applied Psychology*, *50*(1), 117-126.

Lievens, F. (2013). Adjusting medical school admission: Assessing interpersonal skills using situational judgement tests. *Medical education, 47*(2), 182-189. https://doi.org/10.1111/medu.12089

Lievens, F. & Burke, E. (2011). Dealing with the threats inherent in unproctored Internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational and Organizational Psychology*, *84*(4), 817-824. https://doi.org/10.1348/096317910X522672

Lievens, F. & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology*, *96*(5), 927-940. https://doi.org/10.1037/a0023496

Lievens, F. & Peeters, H. (2008). Impact of elaboration on responding to situational judgment test items. *International Journal of Selection and Assessment*, *16*(4), 345-355. https://doi.org/10.1111/j.1468-2389.2008.00440.x

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*(5), 1181-1188. https://doi.org/10.1037/0021-9010.91.5.1181

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*(2), 460-468. https://doi.org/10.1037/a0025741

MacCann, C. & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: theory and data. *Emotion, 8*(4), 540-551. https://doi.org/10.1037/a0012746

Marquaridt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics, 12*(3), 591-612. https://doi.org/10.1080/00401706.1970.10488699

McDaniel, M. A., Hartman, N. S., Whetzel, D. L. & Grubb III, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel psychology, 60*(1), 63-91. https://doi.org/10.1111/j.1744-6570.2007.00065.x

McDaniel, M. A. & Nguyen, N. T. (2001). Situational judgment jests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*(1/2), 103-113. https://doi.org/10.1111/1468-2389.00167

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P. & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology, 96*(2), 327-336. https://doi.org/10.1037/a0021983

McDaniel, M. A., Whetzel, D. L. & Nguyen, N. T. (2006). *Situational judgment tests in personnel selection: A monograph for the International Personnel Management Association Assessment Council. Alexandria*, Alexandra, VA: International Personnel Management Assessment Council.

Motowidlo, S. J. (2003). Job performance. In W.C. Borman, D. R. Ilgen & R. J. Klimoski (Hrsg.), *Handbook of psychology: Industrial and organizational psychology* (S. 39-53). Hoboken, New Jersey: John Wiley & Sons, Inc.

Motowidlo, S. & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology, 95*(2), 321-333. https://doi.org/10.1037/a0017975

Muck, P. M. (2013). Entwicklung von Situational Judgment Tests. *Zeitschrift für Arbeits- und Organisationspsychologie, 57*(4), 185-205. https://doi.org/10.1026/0932-4089/a000125

Mumford, M. D. (1999). Construct validity and background data: Issues, abuses, and future directions. *Human Resource Management Review*, *9*(2), 117-145. https://doi.org/10.1016/S1053-4822(99)00015-7

Mumford, T. V., Van Iddekinge, C. H., Morgeson, F. P. & Campion, M. A. (2008). The Team Role Test: Development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology, 93*(2), 250–267. https://doi.org/10.1037/0021-9010.93.2.250

Murphy, K. R. & Kroeker, L. P. (1988). Dimensions of job performance. In R. Dillon & J. Pellingrino (Hrsg.), *Testing: Applied and theoretical perspectives* (218-247). New York: Praeger.

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*(4), 250-260. https://doi.org/10.1111/j.1468-2389.2005.00322.x

Nikolaou, I. (2014). Social networking web sites in job search and employee recruitment. *International Journal of Selection and Assessment, 22*(2), 179-189. https://doi.org/10.1111/ijsa.12067

Nikolaou, I. (2021). What is the Role of Technology in Recruitment and Selection?. *The Spanish Journal of Psychology, 24*, 1-6. https://doi.org/10.1017/SJP.2021.6

Nixdorf, C. P. (2020). *Handlungskompetenz erfassen mit der Critical Incident Technique: Eine qualitative Forschungstechnik*. Norderstedt: BoD - Books on Demand. https://doi.org/10.25656/01:19375

Okolie, U. C. & Irabor, I. E. (2017). E-recruitment: practices, opportunities and challenges. *European journal of business and management, 9*(11), 116-122.

Ott, M., Ulfert, A. S. & Kersting, M. (2017). „Online-Assessments“ und „Self-Assessments“ in der Eignungsdiagnostik. In D. E. Krause (Hrsg.), *Personalauswahl* (S. 215-242). Wiesbaden: Springer Gabler. https://doi.org/10.1007/978-3-658-14567-5

Petschar, S. & Zavrel, J. (2016). Candidate Experience im E-Recruiting. In T. Verhoeven (Hrsg.), *Candidate Experience* (S. 91-107). Wiesbaden: Springer Gabler. https://doi.org/10.1007/978-3-658-08896-5_9

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), 1-16. https://doi.org/10.1111/1468-2389.00222

Ployhart, R. E. & MacKenzie, W. I., Jr. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Hrsg), *APA handbook of industrial and organizational psychology* (S. 237-252). Washington, DC: American Psychological Association. https://doi.org/10.1037/12170-008

Preuss, P. & Kauffeld, S. (2019). Visualisierung in der Lehre. In S. Kauffeld & J. Othmer (Hrsg.), *Handbuch Innovative Lehre* (S.404-408). Wiesbaden: Springer.

Rammstedt, B., Koch, K., Borg, I. & Reitz, T. (2004). Entwicklung und Validierung einer Kurzskala für die Messung der Big-Five-Persönlichkeitsdimensionen in Umfragen. *ZUMA Nachrichten, 28*(55), 5-28.

Reinecke, J. (2022). Grundlagen der standardisierten Befragung. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S. 949-967). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-37985-8_62

Rogers, S. (2019). How VR, AR and MR are making a positive impact on enterprise*. Forbes Online.* Verfügbar unter: https://www.forbes.com/sites/solrogers/2019/05/09/how-vr-ar-and-mr-are-making-a-positive-impact-on-enterprise/?sh=e0a5535253fb

Rudolf, M. & Müller, J. (2004). *Multivariate Verfahren. Eine praxisorientierte Einführung mit Anwendungsbeispielen in SPSS*. Göttingen: Hogrefe.

Ryf, S., Siegenthaler, P., Fasnacht, D. & Fichter, C. (2022). *NZZ-KMU-Barometer 2022: Lieferkettenprobleme und Fachkräftemangel – die Zukunftsaussichten von Schweizer Unternehmen verdüstern sich.* Zürich: Kalaidos Fachhochschule.

Schäpers, P., Mussel, P., Lievens, F., König, C. J., Freudenstein, J. P. & Krumm, S. (2020). The role of situations in situational judgment tests: Effects on construct saturation, predictive validity, and applicant perceptions*. Journal of Applied Psychology*, *105*(8), 800-818. https://doi.org/10.1037/apl0000457

Schermelleh-Engel, K. & Schweizer, K. (2003). Diskriminante Validität. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 103-110). Weinheim: Beltz.

Scherp, E. (2010). ISK - Inventar Sozialer Kompetenzen von Kanning (2009). *Zeitschrift für Arbeits- und Organisationspsychologie A&O, 54*(4), 193-200. https://doi.org/10.1026/0932-4089/a000030

Schnabel, D., Kelava, A., Seifert, L. & Kuhlbrodt, B. (2014). Konstruktion und Validierung eines multimethodalen berufsbezogenen Tests zur Messung interkultureller Kompetenz. *Diagnostica*, *61*(1), 3-21. https://doi.org/10.1026/0012-1924/a000110

Schneider, G. & Ruoff, G. (2010). Quantitative Methoden. In C. Masala, F. Sauer & A. Wilhelm (Hrsg.), *Handbuch der Internationalen Politik* (S. 236-244). Wiesbaden: VS Verlag für Sozialwissenschaften.

Schreier, M. (2020). Fallauswahl. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 19-39). Wiesbaden: Springer.

Schulz, M., Mack, B. & Renn, O. (2012). *Fokusgruppen in der empirischen Sozialwissenschaft: Von der Konzeption bis zur Auswertung*. Stuttgart: Springer-Verlag.

Serrat, O. (2017). The critical incident technique. In O. Serrat (Hrsg.), *Knowledge solutions* (S. 1077-1083). Singapore: Springer. https://doi.org/10.1007/978-981-10-0983-9_123

Singh, S. P. (2012). Gamification: A strategic tool for organizational effectiveness. *International Journal of Management, 1*(1), 108-113.

Sisson, L. G. & Adams, A. R. (2013). Essential hospitality management competencies: The importance of soft skills. *Journal of Hospitality & Tourism Education, 25*(3), 131-145. https://doi.org/10.1080/10963758.2013.826975

Swisscom (2022). Bereit für deinen Job. Verfügbar unter: https://www.swisscom.ch/de/about/karriere/offene-stellen.html

Talent Board (2021). EMEA Candidate experience research report. Verfügbar unter: https://www.thetalentboard.org/benchmark-research/cande-research-reports/

The International Test Commission (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing, 6*(2), 143-171. https://doi.org/10.1207/s15327574ijt0602_4

Tivian (2022). Experience-Management-Software. Verfügbar unter https://www.tivian.com/de/

Truxillo, D. M., Bauer, T. N., McCarthy, J. M., Anderson, N. & Ahmed, S. M. (2018). Applicant perspectives on employee selection systems. In D. S. Ones, N. Anderson, C. Viswesvaran & H. K. Sinangil (Hrsg.), *The handbook of industrial, work & organizational psychology* (S. 508–532). Thousand Oaks, CA: Sage Reference.

Unity® (2022). Echtzeit-3D-Werkzeuge und mehr. Verfügbar unter https://unity.com/de

Universität Zürich & Fachhochschule Nordwestschweiz (2022). Online-Self-Assessment Psychologie. Verfügbar unter: https://www.psychologie-self-assessment.ch/

Urban, D. & Mayerl, J. (2018). *Angewandte Regressionsanalyse: Theorie, Technik und Praxis*. Wiesbaden: Springer VS.

Verhoeven, T. (2016). *Candidate experience – Ansätze für eine positiv erlebte Arbeitgebermarke im Bewerbungsprozess und darüber hinaus*. Wiesbaden: Springer Gabler.

Verhoeven, T. (2020). Digitale Candidate Experience. In T. Verhoeven (Hrsg.), *Digitalisierung im Recruiting* (S. 51-66). Wiesbaden: Springer Gabler. https://doi.org/10.1007/978-3-658-25885-6_5

Webster, R. (2016). Declarative knowledge acquisition in immersive virtual learning environments. *Interactive Learning Environments, 24*(6), 1319-1333. https://doi.org/10.1080/10494820.2014.994533

Webster, E. S., Paton, L. W., Crampton, P. E. & Tiffin, P. A. (2020). Situational judgement test validity for selection: A systematic review and meta-analysis. *Medical Education, 54*(10), 888-902. https://doi.org/10.1111/medu.14201

Weekley, J. A., Hawkes, B., Guenole, N. & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior, 2*, 295-322. *https://doi.org/10.1146/annurev-orgpsych-032414-111304*

Weekley, J. A. & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*(1), 25-49. https://doi.org/10.1111/j.1744-6570.1997.tb00899.x

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* (157-182). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Whetzel, D. L. & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*(3), 188-202. https://doi.org/10.1016/j.hrmr.2009.03.007

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C. & Anderson, N. R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology, 29*(1), 64-77. https://doi.org/10.1080/1359432X.2019.1681401

Abbildungsverzeichnis

Tabellenverzeichnis