



Bergische Universität Wuppertal

Fakultät für Human- und Sozialwissenschaften, Psychologie

Methodenlehre und Psychologische Diagnostik, Prof. Dr. R. Schulze

**Die Erfassung von Emotional Understanding mit dem  
Empathic Agent Paradigma**

Dissertation zur Erlangung des akademischen Grades Doktor der Philosophie (Dr. phil.) an  
der Fakultät für Human- und Sozialwissenschaften der Bergischen Universität Wuppertal.

Angenommen am 07. Januar 2016.

Dissertation verfasst von

Susan Hellwig

Die Dissertation kann wie folgt zitiert werden:

urn:nbn:de:hbz:468-20160209-142120-0

[<http://nbn-resolving.de/urn/resolver.pl?urn=urn%3Anbn%3Ade%3Ahbz%3A468-20160209-142120-0>]

## Danksagung

Ich möchte mich herzlich bei allen Personen bedanken, die mich in den letzten Jahren auf unterschiedliche Weise bei meiner Arbeit unterstützt haben, sei es in Form der Ermöglichung von Datenerhebungen und/oder durch ein augenscheinlich hohes Maß an Emotional Understanding: Prof. Dr. Philipp Y. Herzberg, Heike Schünemann, Lisa Steffler, Katharina Komorowski, Rebekka Peller, Nils Vahle, Andrea Lautenschütz, Dilek Soysal, Gesa Bintz, Nicole Stüllein, Gronkh, Rainer Moritz, Sarah Moritz sowie Hildegard und Hans-Werner Scheweling. Zudem möchte ich mich bei Philipp Engelberg für die vielen Arbeitstreffen zu jeder Tages- und Nachtzeit bedanken, die nicht nur interessante Diskussionen ermöglicht haben, sondern auch die Freude an der Arbeit verstärkt haben.

Ganz besonderer Dank gilt Prof. Dr. Ralf Schulze, der mich zu einer Arbeit zum Empathic Agent Paradigma angeregt hat und sie in allen Schritten hilfreich begleitet hat. Die ermutigenden Rückmeldungen, anregenden Gespräche sowie die Möglichkeit, regelmäßig das Kolloquium zu nutzen, waren eine sehr wertvolle Unterstützung.

Abschließend möchte ich mich bei meinen Eltern, Erika und Norbert Hellwig, bedanken, die mich zu Lebzeiten und weit darüber hinaus auf diesem Weg maßgeblich unterstützt haben. Ihnen ist diese Arbeit gewidmet.

## Inhaltsverzeichnis

Danksagung .....	2
Abstract .....	8
Kurzzusammenfassung.....	9
1. Einleitung .....	11
2. Theorie .....	21
2.1 Definition von EU .....	21
2.2 Fähigkeitsbasierte Modelle der EI.....	22
2.1.1 Das 1990-Modell der EI.....	22
2.1.2 Das Four-Branch Modell der EI.....	25
2.1.3 Struktur des Four-Branch Modells .....	38
2.3 Abgrenzung zu Mixed Models und Trait EI .....	43
2.3.1 Maximales versus Typisches Verhalten .....	43
2.3.2 Mixed Models .....	44
2.3.3 Trait EI .....	45
2.3.4 Kritik an Mixed Models und Trait-EI .....	46
2.4 EI und Theorien der Intelligenz.....	46
2.4.1 Spearmans $g$ .....	47
2.4.2 Gf-Gc Theorie .....	48
2.4.3 Facetten-Modelle der Intelligenz.....	49
2.4.4 Soziale Intelligenz .....	51
2.4.5 Gardners Theorie der Multiplen Intelligenzen .....	53

2.5 EI und Persönlichkeitseigenschaften.....	54
2.6 EU außerhalb der Theorie der EI .....	56
2.6.1 Affective Forecasting-Fähigkeit.....	56
2.6.2 Empathie.....	59
2.7 Theorien über den Emotionsauslösungsprozess.....	65
2.7.1 Klassische Emotionstheorien .....	66
2.7.2 Appraisal-Theorien.....	67
2.8 Die Messung von EI und EU.....	71
2.8.1 Situational Judgment Tests.....	73
2.8.2 Scoring .....	76
2.8.3 Leistungstests zur Erfassung von EI und EU .....	83
2.8.4 Selbstberichte der EI .....	98
2.9 Optimierung der Messung von EU mit dem Empathic Agent Paradigma .....	100
3. Studie 1.....	104
3.1 Hintergrund und Ziele .....	104
3.2 Methode.....	106
3.2.1 Entwicklung des EAP-Tests.....	106
3.2.2 Entwicklung des Theory Based Test of Emotional Understanding (TBEU) .....	112
3.2.3 Umgang mit Testlet-Effekten.....	113
3.2.4 Stichprobe.....	116
3.2.5 Messinstrumente.....	116
3.2.6 Durchführung .....	119

3.3 Ergebnisse .....	121
3.3.1 Itemselektion des EAP-Tests .....	121
3.3.2 Itemselektion des TBEU .....	122
3.3.3 Testlet-Effekte .....	123
3.3.4 Deskriptive Statistiken der verwendeten Variablen .....	125
3.3.5 Strukturprüfung des EAP-Tests und TBEU .....	126
3.3.6 Faktorkorrelationen zwischen EU, Intelligenz und Persönlichkeitseigenschaften	128
3.4 Diskussion .....	139
4. Studie 2.....	144
4.1 Hintergrund und Ziele .....	144
4.2 Methode.....	145
4.2.1 Stichprobe.....	145
4.2.2 Messinstrumente.....	146
4.2.3 Experimentelles Design.....	147
4.2.4 Klassifikation der Items .....	147
4.2.5 Durchführung .....	148
4.2.6 Auswertung .....	149
4.3 Ergebnisse .....	150
4.3.1 Deskriptive Statistiken und Reliabilitätsschätzungen für die eingesetzten Testverfahren.....	150
4.3.2 Effekte auf die Testleistung.....	151
4.3.3 Kontrolle von Gruppenunterschieden hinsichtlich der EU-Fähigkeit.....	153

4.3.4 Zusammenhang der empirischen mit den vorhergesagten Itemschwierigkeiten...	154
4.4 Diskussion .....	155
5. Studie 3.....	158
5.1 Hintergrund und Ziele .....	158
5.2 Methode.....	164
5.2.1 Entwicklung des DEF-Tests .....	164
5.2.2 Maße der Akkuratheit im DEF-Test.....	166
5.2.3 Analysen dyadischer Daten .....	167
5.2.4 Stichprobe.....	171
5.2.5 Messinstrumente.....	172
5.2.6 Durchführung .....	173
5.2.7 Auswertung .....	174
5.3 Ergebnisse.....	174
5.3.1 Itemanalysen des EAP-Tests und DEF-Tests.....	174
5.3.1 Deskriptive Statistiken und Reliabilitätsschätzungen .....	176
5.3.2 Strukturprüfung des DEF-Tests: dyadische CFA.....	178
5.3.4 Actor-Partner Interdependence Modell .....	181
5.3.5 Zusammenhänge mit verbaler fluider Intelligenz und den Big-Five Persönlichkeitsfaktoren .....	183
5.4 Diskussion .....	186
6. Diskussion .....	189
6.1 Zusammenfassende Bewertung des EAP-Tests .....	189

6.2 Zukünftige Studien zum EAP-Test .....	195
6.3 Weitere Anwendungen für das EAP .....	197
6.4 Fortschritt im Bereich EI.....	198
Zusammenfassung.....	200
Referenzen.....	203

### **Elektronischer Anhang**

Appendix A: Itemstatistiken für den EAP-Test

Appendix B: Itemstatistiken für den TBEU

Appendix C: Faktorkorrelationen Parcelvariante 2

Appendix D: Faktorkorrelationen Parcelvariante 3



## Abstract

Emotional Understanding (EU) constitutes an ability construct that is related to Emotional Intelligence (EI). Conceptual ambiguities as well as a lack of a veridical scoring criterion for the evaluation of test responses are two of the most serious problems in this domain. To improve the measurement of EU, Schulze and Roberts recently developed the Empathic Agent Paradigm (EAP; Schulze, Holtzman, MacCann & Roberts, 2009). Tests relying on the EAP are characterized by a two phases design: During the first phase (the acquisition phase) the test-taker observes the emotional reactions of one or more target persons in a given situation. In the second phase (the application phase) the test-taker has to correctly apply the acquired knowledge about the target persons in a comparable situation. In this thesis, a new computer-based maximum performance measure of EU was developed using the EAP: the EAP-Test. In three studies, the psychometric quality and evidence for the validity of the EAP-Test were examined. For the EAP-Test acceptable reliability estimates were found on testlet level ( $\alpha = .70$  to  $.78$ ). In Study 1 ( $N = 210$ ), correlational validity evidence was provided on latent variable level with measures of classical intelligence, the Big-Five personality factors, and a newly developed theory-based test of EU. In Study 2, experimental validity evidence was examined for the EAP-Test. Participants ( $N = 158$ ) were randomly assigned to two testing conditions: The EAP-Test was either presented excluding (experimental group) or including the acquisition phase (control group). As expected, the experimental group scored lower on the EAP-Test and this effect was higher for items presenting target persons with uncommon emotional reactions. Results support the assumption that the acquisition phase plays an important role in the response process. In a third study the EAP-Test showed a positive relationship with a newly developed measure of empathic accuracy in a dyadic sample of 146 couples. Overall, the EAP-Test provides a promising new approach to the assessment of EU.

## Kurzzusammenfassung

Emotional Understanding (EU) stellt ein Fähigkeits-Konstrukt dar, das mit Emotionaler Intelligenz (EI) verbunden ist. Konzeptuelle Unklarheiten sowie das Fehlen eines veridikalen Scoring-Kriteriums sind zwei der schwerwiegendsten Probleme in dieser Domäne. Mit dem Ziel der Verbesserung der Messung von EU, entwickelten Schulze und Roberts das Empathic Agent Paradigma (EAP; Schulze, Holtzman, MacCann & Roberts, 2009). Tests, die auf dem EAP beruhen, sind durch ein Zwei-Phasen-Design charakterisiert: In der ersten Phase (der Aneignungsphase) beobachtet die Testperson die emotionalen Reaktionen einer oder mehrerer Zielpersonen in einer vorgegebenen Situation. In der zweiten Phase (der Aneignungsphase) muss die Testperson das erworbene Wissen über die Zielperson in einer vergleichbaren Situation korrekt anwenden. In dieser Arbeit wurde unter Verwendung des EAP ein neuer, computerbasierter Test maximalen Verhaltens für EU entwickelt: der EAP-Test. In drei Studien wurden die psychometrische Qualität sowie Belege für die Validität des EAP-Tests untersucht. Für den EAP-Test wurden akzeptable Reliabilitätsschätzungen auf Ebene von Testlets gefunden ( $\alpha = .70$  bis  $.78$ ). In Studie 1 ( $N = 210$ ) wurden korrelative Validitäts-Belege auf Ebene latenter Variablen mit Maßen für klassische Intelligenz, für die Big-Five Persönlichkeitsfaktoren sowie mit einem neu entwickelten theorie-basierten Test für EU geliefert. In Studie 2 wurde experimentelle Validitätsevidenz für den EAP-Test erbracht. Die Teilnehmerinnen und Teilnehmer ( $N = 158$ ) wurden zufällig zwei Test-Bedingungen zugewiesen: Der EAP-Test wurden entweder ohne die (Experimentalgruppe) oder mit der Aneignungsphase (Kontrollgruppe) präsentiert. Wie erwartet erzielte die Experimentalgruppe geringere Testwerte, wobei der Effekt für Items höher war, in denen die Zielpersonen ungewöhnliche emotionale Reaktionen aufwiesen. Die Ergebnisse unterstützen die Annahme, dass die Aneignungsphase eine wichtige Rolle im Antwortprozess spielt. In der dritten Studie wies der EAP-Test eine positive Beziehung zu einem neu entwickelten Maß für Empathic

Accuracy in einer dyadischen Stichprobe von 146 Paaren auf. Insgesamt bietet der EAP-Test einen vielversprechenden neuen Ansatz zur Messung von EU.

## 1. Einleitung

*Emotionale Intelligenz* (EI) wurde als eine neue Form der Intelligenz vorgeschlagen, die vier Fähigkeitskonstrukte umfasst (Mayer & Salovey, 1997). *Emotional Understanding* (EU) stellt eine dieser Fähigkeiten dar. Im Vergleich zu etablierten Theorien der Intelligenz, ist EI ein relativ junges Konstrukt. Eine erste Definition der EI wurde von Salovey und Mayer im Jahre 1990 vorgeschlagen, zusammen mit einem vorläufigen Modell der EI. Sieben Jahre später schlugen die Autoren das sogenannte *Four-Branch Modell der EI* vor, das auf Grundlage des ersten Modells entwickelt wurde (Mayer & Salovey, 1997). Obwohl der Begriff EI bereits in einigen früheren Publikationen verwendet wurde (z. B. Leuner, 1966), gilt der Artikel von Salovey und Mayer (1990) als Startpunkt der Forschung zur EI, da die beiden Autoren hier den ersten Versuch unternahmen, den Begriff EI zu konzeptualisieren.

Obwohl EI ein relativ junges Konstrukt ist, wurden bis zum heutigen Zeitpunkt zahlreiche Forschungsarbeiten hierzu angeregt. Das wissenschaftliche Interesse an EI wird durch eine hohe und kontinuierlich zunehmende Anzahl von Publikationen zu diesem Thema reflektiert. Um diese Sachlage zu veranschaulichen, wird in Abbildung 1.1 die Anzahl von Treffern pro Jahr für die Suchanfrage „emotional intelligen\*“ in der Datenbank PsycINFO von 1990 bis 2014 präsentiert. Wie aus Abbildung 1.1 entnommen werden kann, stieg die Anzahl neuer Veröffentlichungen pro Jahr zwischen der Mitte der 1990er bis zur Mitte der 2000er rapide an. Bis heute verbleibt die Anzahl neuer Publikationen pro Jahr auf einem hohen Niveau, was das stetige Interesse zu diesem Thema verdeutlicht.

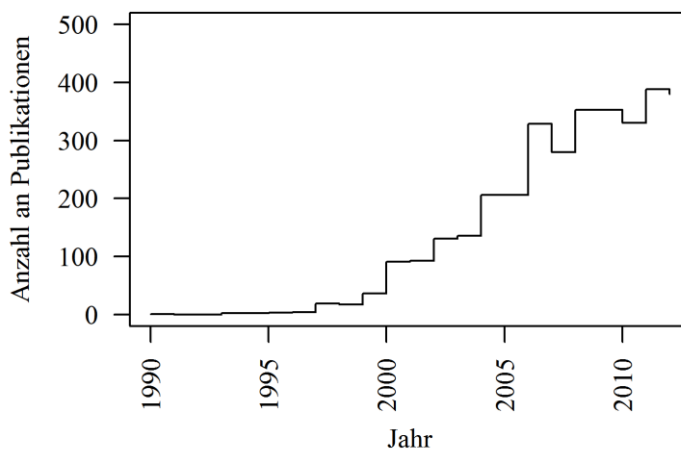


Abbildung 1.1. Anzahl der Treffer für die Suchanfrage „emotional intelligenz\*“ in der Datenbank PsycINFO von 1990 bis 2014.

Im Vergleich zu anderen Intelligenz-Konstrukten, hat EI auch in den Medien große Aufmerksamkeit erhalten. Seit Einführung des Konstruktes wurden eine Reihe von populärwissenschaftlichen Büchern, Zeitungsartikeln und Dokumentarfilmen veröffentlicht. EI wurde sogar im Bereich der Musik thematisiert (*The Emotional Intelligence Song* von Kristopher Stone), was ein ungewöhnlicher Umstand für ein Fähigkeitskonstrukt ist. Die Popularität von EI, im Vergleich zu anderen Intelligenz-Konstrukten, scheint in einigen Behauptungen über das Konstrukt begründet zu liegen, die für eine Form der Intelligenz erstaunlich und attraktiv erscheinen. Diese Ideen wurden hauptsächlich durch den Bestseller „*Emotional Intelligence: Why it can Matter More Than IQ*“ von Daniel Goleman (1995) verbreitet sowie durch einen *Times Magazine* Artikel, der im gleichen Jahr folgte (Matthews, Zeidner & Roberts, 2002). Goleman (1995) behauptete, dass EI ein wichtiger Prädiktor für ein erfolgreiches Leben ist und in einigen Domänen sogar eine bessere Vorhersagekraft aufweist, als etablierte Formen der Intelligenz. Darüber hinaus vertrat Goleman die Ansicht, dass EI

weniger durch Gene determiniert ist, als anderen Arten der Intelligenz und somit durch Lernerfahrungen substantiell verbessert werden kann (Goleman, 1995). Im Einklang mit der Vorstellung, dass sich die Fähigkeit der EI erhöhen lässt, sind heute viele Angebote zur Verbesserung von EI zu finden, wie bspw. Schulungen, Selbsthilfebücher oder Lernspielzeuge. Golemans Buch hat die Wahrnehmung eines breiten Publikums auf EI stark beeinflusst. Ein Grund für die Überzeugungskraft seines Buches könnte in der Art liegen, wie die darin getroffenen Behauptungen untermauert wurden. Goleman fasste in seinem Buch diverse Befunde aus unterschiedlichen Feldern der Psychologie narrativ zusammen und reicherte diese mit zahlreichen Anekdoten an, die scheinbar seine Behauptungen belegen. In der wissenschaftlichen Literatur wurden Golemans Behauptungen aufgrund des Fehlens einer empirischen Basis oft kritisiert (z. B. Matthews et al., 2002). Zwar mag Golemans Bestseller auch zu Recht zugeschrieben werden, für die hohe Popularität von EI und damit zum Teil auch für den Schub an wissenschaftlichen Veröffentlichungen beigetragen zu haben (Neubauer & Freudenthaler, 2005); allerdings haben die darin getroffenen Behauptungen zu einigen konzeptuellen Unklarheiten in der wissenschaftlichen Forschung zur EI geführt, die sich bis heute auswirken (siehe Mayer, Salovey & Caruso, 2000).

In der wissenschaftlichen Literatur wird EI ebenfalls als ein wichtiger Prädiktor in verschiedenen Domänen diskutiert, wie im Arbeitskontext (z. B. Chien Farh, Seo & Tesluk, 2012; Joseph & Newman, 2010), im pädagogischen Kontext (z. B. Lopes, Mestre, Guil, Kremenitzer & Salovey, 2012; Rivers, Brackett, Reyes, Mayer, Caruso & Salovey, 2012), für die physische Gesundheit (z. B. Keefer, Parker & Saklofske, 2009; Schutte, Malouff, Thorsteinsson, Bhullar & Rooke, 2007; Rossen & Kranzler, 2009) oder im Rahmen der klinischen Psychologie (Davis & Humphrey, 2012; Kee et al., 2009).

Obwohl das Interesse an EI und die Bemühungen das Konstrukt zu etablieren hoch sind, ist der Forschungsfortschritt in dieser Domäne interessanterweise nicht so weit, wie man

erwarten könnte. Bis heute beschäftigen sich viele Publikationen immer noch mit zwei fundamentalen Fragen: Existiert eine Fähigkeit der EI? Wie kann EI gemessen werden? Beide Fragen sind eng miteinander verbunden, da Schlussfolgerungen über EI, die auf empirischer Evidenz beruhen, einen Ansatz zur Operationalisierung des Konstruktes benötigen. Trotz des Umfangs an Publikationen und einigen vielversprechenden Ansätzen in diesem Feld (z. B. MacCann & Roberts, 2008), können beide Fragen noch nicht definitiv beantwortet werden. Tatsächlich sind Forscherinnen und Forscher in dieser Domäne noch weit von einer gemeinsamen Sichtweise auf diese beiden Fragen entfernt.

In Bezug auf die Existenz einer Fähigkeit der EI, werteten einige Forscherinnen und Forscher die Suche nach EI als erfolglose Bemühung (z. B. Landy, 2006; Schulte, Ree & Carretta, 2004), während andere die Belege für die Existenz einer neuen Intelligenzform als vielversprechend (Chiarrochi, Chan & Caputi, 2000) oder sogar als ausreichend betrachteten (Mayer, Caruso & Salovey, 1999). Insbesondere gibt es unterschiedliche Schlüsse darüber, ob EI trennbar von etablierten Formen der Intelligenz oder von Persönlichkeitseigenschaften ist. Einige dieser inkonsistenten Schlussfolgerungen über die Existenz von EI scheinen auf schwerwiegenden konzeptuellen Ambiguitäten zu basieren. Der Begriff EI wurde für sehr unterschiedliche Konzepte aus der Fähigkeits- und Persönlichkeitsdomäne verwendet (für eine Übersicht siehe Mayer et al., 2000). Während einige EI auf ein Set mentaler Fähigkeiten begrenzten (Mayer & Salovey, 1997), betrachten andere EI als eine Mischung aus verschiedenen dispositionellen Konzepten (d. h. Verhaltensstile, Motivationen oder Interessen) und Fähigkeiten (z. B. Bar-On, 2006; Goleman, 1995; siehe auch Mayer et al., 2000). Beispiele für solche Dispositionen, die unter EI gefasst wurden, sind Impulsivität, Trait-Optimismus, Selbstwert oder Selbst-Motivation (Petrides, Furnham & Mavroveli, 2007). Angesichts dieser Heterogenität von Konzeptualisierungen der EI überrascht es nicht, dass Befunde und Schlussfolgerungen über EI teilweise stark voneinander abweichen.

Um die Unterschiedlichkeit der Modelle berücksichtigen zu können, wurden Vorschläge für Taxonomien gemacht, wie die Unterscheidung von *Ability-* und *Trait-EI* sowie *Ability Models* und *Mixed Models* (Mayer et al., 2000; Petrides & Furnham, 2000). Zusätzlich gab es Anstrengungen, potentiell valide von invaliden Konzeptionen der EI abzugrenzen. Beispielsweise argumentierten Schulze, Wilhelm und Kyllonen (2007), dass der Begriff *Intelligenz* angesichts semantischer, theoretischer sowie empirischer Gründe ausschließlich Fähigkeitskonstrukten vorbehalten sein sollte.

Das einzige Modell, in dem EI strikt auf die Fähigkeitsdomäne begrenzt ist, ist das Four-Branch Modell der EI (Mayer & Salovey, 1997). Doch sogar innerhalb der Forschung zu Operationalisierungen des Four-Branch Modells besteht Uneinigkeit darüber, ob EI tatsächlich als neue Intelligenzform betrachtet werden kann (Brody, 2004; Schulte et al., 2004) und ob sich die vorgeschlagene faktorielle Struktur der EI empirisch zeigen lässt (z. B. Maul, 2012). Um Theorien der EI prüfen zu können, ist die Verfügbarkeit eines Ansatzes zur Messung notwendig. Schulze et al. (2007) wiesen darauf hin, dass empirische Bewertungen der Theorien über EI auch von der Qualität der verfügbaren Messinstrumente abhängen (Schulze et al., 2007).

Unglücklicherweise leiden die aktuell verfügbaren Messansätze unter einigen schwerwiegenden Problemen. Hierbei wurden mindestens zwei Hindernisse identifiziert, die das Voranschreiten des Forschungsprozesses aufhalten. Das erste Problem bezieht sich auf einen Mangel an Ansätzen zur Erfassung von EI. Die Messung von EI und ihrer Komponenten macht die Verfügbarkeit von Leistungstests notwendig. Leistungstests sind durch Items gekennzeichnet, die das maximale Verhalten einer Testperson provozieren (Cronbach, 1990). Die Fähigkeit wird dabei durch die gezeigte Leistung direkt messbar gemacht. Untersuchungen von EI auf Grundlage von Leistungstests werden derzeit durch einen einzelnen Messansatz dominiert: den Mayer-Salovey-Caruso Emotional Intelligence



Test (MSCEIT; Mayer, Salovey & Caruso, 2002) sowie seinen Vorgänger, der Multifactor Emotional Intelligence Scale (MEIS; Mayer et al., 1999). Neben den wenigen Leistungstests existiert eine große Anzahl von Selbstberichtsverfahren, die der Messung von EI dienen sollen. Diese Instrumente sind fast ausschließlich mit dispositionellen Konzeptualisierungen von EI assoziiert. Selbstberichte der EI beinhalten Fragen über das typische Verhalten der Getesteten oder Selbsteinschätzungen von emotionsbezogenen Fähigkeiten. Beispiele für Selbstberichtsverfahren zur Messung von EI sind das Emotional Quotient Inventory (EQ-i; Bar-On, Brown, Kirkcaldy & Thomé, 2000) die Schutte Emotional Intelligence Scale (SEIS; Schutte, Malouff, Hall, Haggerty, Cooper, Golden & Dornheim, 1998), der Trait Emotional Intelligence Questionnaire (TEIQue; Petrides, Pita & Kokkinaki, 2007) sowie die Wong & Law Emotional Intelligence Scale (WLEIS; Wong & Law, 2002). Dabei steigt die bereits hohe Anzahl von Selbstberichtsverfahren stetig an. Selbstberichte wurden von vielen Forscherinnen und Forschern als unangemessene Ansätze zur Messung von Fähigkeitskonstrukten bewertet (Schulze et al., 2007). Einige Belege sprechen dafür, dass sich Fähigkeiten nur ungenügend über Selbstberichte abbilden lassen (z. B. Freund & Kasten, 2012; Lievens, Klehe & Libbrecht, 2011). Da nur wenige Leistungstests existieren und der MSCEIT/MEIS-Ansatz die Messung von EI dominiert, beruhen fast alle Erkenntnisse über EI auf einem einzelnen Messinstrument, was als unerwünschter Umstand identifiziert wurde (Schulze et al., 2007; MacCann & Roberts, 2008).

Ein zweites Problem in der Erforschung von EI bezieht sich auf Zweifel, die über die Angemessenheit der aktuell verwendeten Scoring-Mechanismen in Leistungstests erhoben wurden (z. B. Roberts, Zeidner & Matthews, 2001). Leistungstests der EI benötigen ein Scoring-Kriterium, anhand dessen korrekte und inkorrekte Testantworten bewertet werden können. In den meisten Intelligenztests beruht das Scoring auf einem regelgebundenen System, das eindeutig determiniert, welche Antworten auf ein Item korrekt oder inkorrekt sind (Roberts et al., 2001). Zum Beispiel stellt für einen Vocabulary-Test die Semantik der

Sprache ein veridikales Kriterium für das Scoring dar (Roberts et al., 2001). Für den Inhaltsbereich der Emotionen konnte ein solcher veridikaler Bewertungsstandard noch nicht identifiziert werden (MacCann, Roberts, Matthews & Zeidner, 2004). Angesichts des Fehlens eines veridikalen Scoring-Kriteriums in der Domäne der EI, werden in den bisherigen Leistungstests, wie dem MSCEIT und MEIS, entweder Gruppen-Urteile von Experten oder Novizen (Experten-Scoring oder Konsensus-Scoring) oder Selbstberichte von Zielpersonen über emotionale Reaktionen (Target-Scoring) herangezogen. Eine seltene Ausnahme stellt der Ansatz von MacCann und Roberts (2008) dar: im Situational Test of Emotional Understanding (STEU) basiert das Scoring auf einer bestimmten Theorie über den Prozess der Emotionsauslösung. Der Mangel an Scoring-Methoden kann ein Grund dafür sein, dass derzeit ein Defizit an verfügbaren Leistungstests verzeichnet werden muss.

Viele Forscherinnen und Forscher stimmten darin überein, dass die Entwicklung neuer Messansätze und neuer Scoring-Rationale eines der wichtigsten Aufgaben in der Forschung zur EI ist (z. B. Farelly & Austin, 2007; MacCann & Roberts, 2008; Orchard, MacCann, Schulze, Matthews, Zeidner & Roberts, 2009; Schulze et al., 2007; Van Rooy, Viswesvaran & Pluta, 2005). Glücklicherweise haben einige Forscherinnen und Forscher neue Ideen vorgeschlagen, die zu einer Verbesserung der Messung von EI beitragen könnten (MacCann & Roberts, 2008; Orchard et al., 2009; Schulze et al., 2009). Mit dem Ziel die Unterschiedlichkeit von Messansätzen zu erhöhen und ein verbessertes Scoring-Rational zu liefern, wurde ein neues Paradigma von Schulze und Roberts eingeführt: das Empathic Agent Paradigma (EAP) (Schulze et al., 2009; Schulze, Hellwig, MacCann & Roberts, 2010). Das EAP wurde als besonders nützlich für die Konstruktion und das Scoring von Tests zur Erfassung von EU betrachtet. Tests, die auf dem EAP basieren, weisen ein Zwei-Phasen-Design auf: Während der ersten Phase (der Aneignungsphase) werden den Getesteten die emotionalen Kontingenzen einer oder mehrerer Zielpersonen präsentiert. Diese Emotionskontingenzen können Verbindung zwischen Ereignissen und Emotionen, Ketten von

Emotionen, oder Handlungen, die emotionalen Zuständen folgen, darstellen (Orchard et al., 2009). In der zweiten Phase (der Anwendungsphase) besteht die Aufgabe der Testpersonen darin, das angeeignete Wissen über die Zielperson in einer neuen, vergleichbaren Situation korrekt anzuwenden. Das Zwei-Phasen-Design erlaubt es dem Testanwender, die Testantworten als eindeutig richtig oder falsch zu bewerten, da die Korrektheit einer Testantwort im Prozess der Konstruktion der Aneignungsphase festgelegt wird. Im Gegensatz zu den verfügbaren Messansätzen, hängt die Angemessenheit des Scoring nicht von der Richtigkeit menschlicher Urteile (d. h. MSCEIT/MEIS-Ansatz) oder einer bestimmten Theorie des Emotionsauslöseprozesses ab (d. h. STEU-Ansatz).

Das Ziel dieser Arbeit bestand darin, zu untersuchen, ob das EAP ein nützliches Werkzeug für die Entwicklung von Tests zur Erfassung von EU ist. Zu diesem Zweck wurde im Rahmen dieser Arbeit ein neuer Leistungstest zur Messung von EU basierend auf dem EAP entwickelt: der Empathic Agent Paradigm-Test (EAP-Test).

Im theoretischen Teil wird zunächst ein konzeptueller Rahmen geliefert, in dem EU definiert und von andersartigen und ähnlichen Konstrukten abgegrenzt wird. Wie erwähnt, können konzeptuelle Unklarheiten auf die Verwendung des Begriffs EI für sehr unterschiedliche Konzepte aus der Fähigkeits- und Persönlichkeitsdomäne zurückgeführt werden. Eine weitere Quelle der Ambiguität liegt in der Tatsache, dass EI und ihre Komponenten, wie EU, selten mit ähnlichen emotionalen Fähigkeiten in Verbindung gebracht wurden, die außerhalb des Forschungsfeldes der EI diskutiert wurden. Einige dieser Konzepte könnten mit EI und EU teilweise überlappen (z. B. Social Intelligence und Social Understanding), während andere sogar sehr ähnlich erscheinen (z. B. Empathic Accuracy, Theory of Mind) und vielleicht sogar identisch mit ihnen sind. Darüber hinaus werden im theoretischen Teil die aktuellen Ansätze zur Messung von EI und EU thematisiert.

Abschließend wird das EAP als potentiell nützliches Werkzeug zur Erfassung von EU diskutiert.

Im empirischen Teil der Thesis wird die Entwicklung des EAP-Tests beschrieben und Evidenz für seine psychometrische Qualität sowie für seine Validität präsentiert. Das Ziel bei der Gestaltung des konkreten Testdesigns bestand darin, den möglichen Einfluss anderer Konstrukte, wie Emotionswahrnehmungsfähigkeit, zu kontrollieren.

Danach werden Ergebnisse aus drei Studien präsentiert, in denen der EAP-Test erprobt wurde. In der ersten Studie wird die angenommene eindimensionale Struktur des EAP-Tests geprüft. Zudem wird in dieser Studie geprüft, ob der EAP-Test, im Sinne konvergenter und diskriminanter Validitätsbelege, die theoretisch angenommenen Beziehungen zu Maßen für klassische Intelligenz, Persönlichkeitseigenschaften und EU aufweist. Aufgrund eines Mangels an verfügbaren Tests für EU, wurde im Rahmen dieser Studie ein weiteres Instrument zur theoriebasierten Messung von EU entwickelt.

In der zweiten Studie werden experimentelle Belege für den EAP-Test erbracht. Hierbei wird die grundlegende Annahme über den Antwortprozess im EAP getestet, dass Testpersonen in der Anwendungsphase tatsächlich auf Informationen über die Zielpersonen aus der Aneignungsphase zurückgreifen, und die Items nicht allein auf Grundlage allgemeinen Wissens über Emotionen beantworten. In diesem Zusammenhang wurde der Experimentalgruppe eine um die Aneignungsphase reduzierte Version des EAP-Tests vorgelegt, während die Kontrollgruppe die vollständige Version erhielt. Falls die Annahme hält, dass Testpersonen ihre Vorhersagen auf Basis der Informationen über die Zielpersonen aus der Aneignungsphase treffen, sollte die Experimentalgruppe geringere Testwerte erzielen. In diesem Zusammenhang wird zudem eine Annahme über ein schwierigkeitsstiftendes Merkmal der Items getestet.

In der dritten Studie wird die Beziehung des EAP-Tests mit der Fähigkeit, die emotionalen Reaktionen einer Partnerin bzw. eines Partner vorhersagen zu können, in einer dyadischen Stichprobe untersucht, um weitere korrelative Validitätsevidenz zu liefern. Zu diesem Zweck wurde in dieser Studie ein dyadisches Instrument zur Erfassung von Empathic Accuracy Fähigkeit entwickelt (siehe Ickes, 2001).

Zusammengefasst wird mit dieser Thesis intendiert, a) einen theoretischen Rahmen für die Entwicklung von Tests für EU zu liefern, b) ein neues Instrument zur Erfassung von EU zu entwickeln, das auf einem vielversprechenden neuen Konstruktionsansatz beruht sowie c) eine empirische Basis zur Bewertung des EAP-Tests sowie des EAP zu liefern.

## 2. Theorie

### 2.1 Definition von EU

EU ist eng mit der Theorie der EI verknüpft. EI wurde von Mayer, Roberts und Barsade (2008) definiert als „the ability to carry out accurate reasoning about emotions and the ability to use emotions and emotional knowledge to enhance thought“ (S. 511). Im Zusammenhang mit EI wurde EU zum ersten Mal im sogenannten Four-Branch Modell der EI von Mayer und Salovey (1997) genannt. EU stellt in diesem Modell eine von vier emotionsbezogenen Fähigkeiten dar, die unter dem Begriff EI zusammengefasst werden. Im Bereich der EI wurde EU auf unterschiedliche Weise definiert und es wurden verschiedene Aspekte darunter zusammengefasst (siehe Mayer & Salovey, 1997; Mayer et al., 2000). Als Kern dieser Konzeptualisierungen wird in dieser Arbeit die folgende Definition verwendet: EU ist die Fähigkeit, die regelhaften Prozesse, die der Auslösung und dem Verlauf von Emotionen zu Grunde liegen, identifizieren zu können und zur korrekten Vorhersage von Emotionen heranziehen zu können. Bezüglich der Definition von Mayer et al. (2008) entspricht EU damit dem Teilaspekt des „accurate reasoning about emotions“ (S. 511).

Eine Voraussetzung für das Verstehen von Emotionen ist die Annahme, dass Emotionen bestimmten Regelhaftigkeiten folgen. Eine Person mit hoher EU Fähigkeit kann die Bedingungen erkennen, unter denen eine Emotion auftritt, sich hinsichtlich ihrer Intensität verändert oder in andere Emotionen übergeht. Auf Basis des Verständnisses dieser Regelhaftigkeiten können Emotionen, gegeben eine auslösende Bedingung, akkurat vorhergesagt werden. Regelhaftigkeiten von Emotionen sind selbst Gegenstand einiger Theorien, die sich mit dem Auslösungsprozess von Emotionen beschäftigen (siehe z. B. Roseman & Smith, 2001). Beispielsweise wird im Rahmen der sogenannten Appraisal-Theorien angenommen, dass spezifische Muster an Bewertungen einer Situation Auslöser von

Emotionen sind. Die Annahmen solcher Theorien wurden bereits als eine Grundlage für die Konstruktion von Tests von EU herangezogen (siehe MacCann & Roberts, 2008).

Zwar stellt EU einen wichtigen Bestandteil der Theorie der EI dar, die Idee einer Fähigkeit, Emotionen verstehen zu können, wurde allerdings auch als Teil von anderen Konzepten außerhalb der EI diskutiert. EU weist eine hohe inhaltliche Überlappung mit dem Konzept der fähigkeitsbasierten Empathie auf. Fähigkeitsbasierte Empathie wurde definiert als das Vermögen, sich in die Emotionen anderer Personen einfühlen zu können oder diese vorhersagen zu können (z. B. Bernhardt & Singer, 2012). Dabei handelt es sich um den Aspekt von EU, der sich auf die Fähigkeit im Umgang mit anderen Personen bezieht. Ein anderes Konzept außerhalb der EI, das ebenfalls eng mit EU verknüpft ist, ist die Affective Forecasting Fähigkeit. Affective Forecasting Fähigkeit bezieht sich auf das Verstehen und die Vorhersage der eigenen Emotionen (z. B. Wilson & Gilbert, 2003). Fähigkeitsbasierte Empathie und Affective Forecasting Fähigkeit wurden bisher eher isoliert von Intelligenztheorien betrachtet. Mit der Theorie der EI wurde versucht, EU sowie weitere emotionsbezogenen Fähigkeiten mit Intelligenztheorien in Verbindung zu setzen.

## **2.2 Fähigkeitsbasierte Modelle der EI**

### **2.1.1 Das 1990-Modell der EI**

Im Jahre 1990 schlugen Salovey und Mayer das Konzept sowie ein erstes Modell der EI vor. Im Rahmen dieses ersten Versuches einer Konzeptualisierung, wurde EI als ein Teil sozialer Intelligenz definiert, der die Fähigkeit „to monitor one’s own and other’s feelings and emotions, to discriminate among them and to use this information to guide one’s thinking and actions“ (Salovey & Mayer, 1990, S. 189) umfasst. Der Ausgangspunkt für den Vorschlag einer neuen Form von Intelligenz lag in der Beobachtung der Autoren, dass es eine Reihe von Studien zu emotionsbezogenen Fähigkeiten gab, denen eine konzeptuelle Verbindung

untereinander fehlte. Forschungsergebnisse zu diesen Fähigkeiten wurden daher ohne Bezug zueinander in unterschiedlichsten Domänen der psychologischen Forschung veröffentlicht (Salovey & Mayer, 1990). Aus Sicht der Autoren konnte der Beitrag dieser Studien zur Erforschung emotionsbezogener Fähigkeiten unter diesen Umständen nur gering sein.

Um den fehlenden theoretischen Rahmen zu entwickeln, synthetisierten Salovey und Mayer (1990) die Erkenntnisse aus Studien über emotionsbezogene Fähigkeiten zu einem ersten Modell der EI. EI wurde von den Autoren als Schnittstelle zwischen den beiden, oft als gegensätzlich betrachteten, Begriffen Emotion und Kognition vorgeschlagen (Mayer et al., 2000; Salovey & Mayer, 1990). Das 1990-Modell der EI wird in Abbildung 2.1 dargestellt. In diesem Modell wurde EI als eine Form von Intelligenz konzeptualisiert, die drei Teilfähigkeiten umfasst: a) die Bewertung und den Ausdruck von eigenen Emotionen und der Emotionen anderer Personen, b) das Regulieren der eigenen Emotionen und der Emotionen anderer Personen und c) die Nutzung von Emotionen beim Lösen von Problemen (Salovey & Mayer, 1990). Den Teilfähigkeiten Bewertung und Ausdruck von Emotionen bei sich selbst und anderen sowie der Nutzung von Emotionen wurden noch weitere Teilfähigkeiten untergeordnet (siehe Salovey & Mayer, 1990).

Der Abbildung kann zudem entnommen werden, dass das Konzept EU in diesem Modell noch nicht enthalten war. Allerdings ist unter der Teilfähigkeit der Bewertung und des Ausdrucks von Emotionen das Konzept der Empathie gefasst, das Salovey und Mayer (1990) als Fähigkeit definierten „die Gefühle des anderen zu verstehen“ (S. 194). Somit ist ein Teilaspekt von EU, der sich auf die Emotionen Anderer bezieht, bereits in diesem Modell enthalten. Aus Sicht der Autoren stellt Empathie einen wichtigen Aspekt der EI dar, der eine positive Wirkung auf Beziehungen, auf die Lebenszufriedenheit und Gesundheit haben kann (Salovey & Mayer, 1990).



Aus den Beschreibungen der vorgeschlagenen Fähigkeiten im Rahmen des 1990-Modells der EI wird eine Schwäche des Modells unmittelbar ersichtlich. Obwohl die Autoren EI als eine Form der Intelligenz auffassten, waren die vorgeschlagenen Elemente des Modells nicht allein auf Fähigkeitskonstrukte beschränkt. Empathie wurde beispielsweise als „Motivator für altruistisches Verhalten“ (Salovey & Mayer, 1990, S. 194) beschrieben, was eher einem Verhaltensstil als einem Fähigkeitskonzept entspricht. Darüber hinaus wurden in diesem Modell weitere Konzepte vorgeschlagen, wie eine „positive Einstellung zum Leben“ (Salovey & Mayer, 1990, S. 200), die nicht in den Bereich der Fähigkeiten zählen. Die Autoren des 1990-Modells der EI kritisierten später selbst, dass die Konstrukte nicht eindeutig auf Fähigkeiten begrenzt waren (siehe Mayer et al., 2000). Trotz dieser gravierenden Mängel, bot das Modell dennoch eine erste Grundlage zur Erforschung von EI und regte in den folgenden Jahren erste Studien in dieser neuen Domäne an.

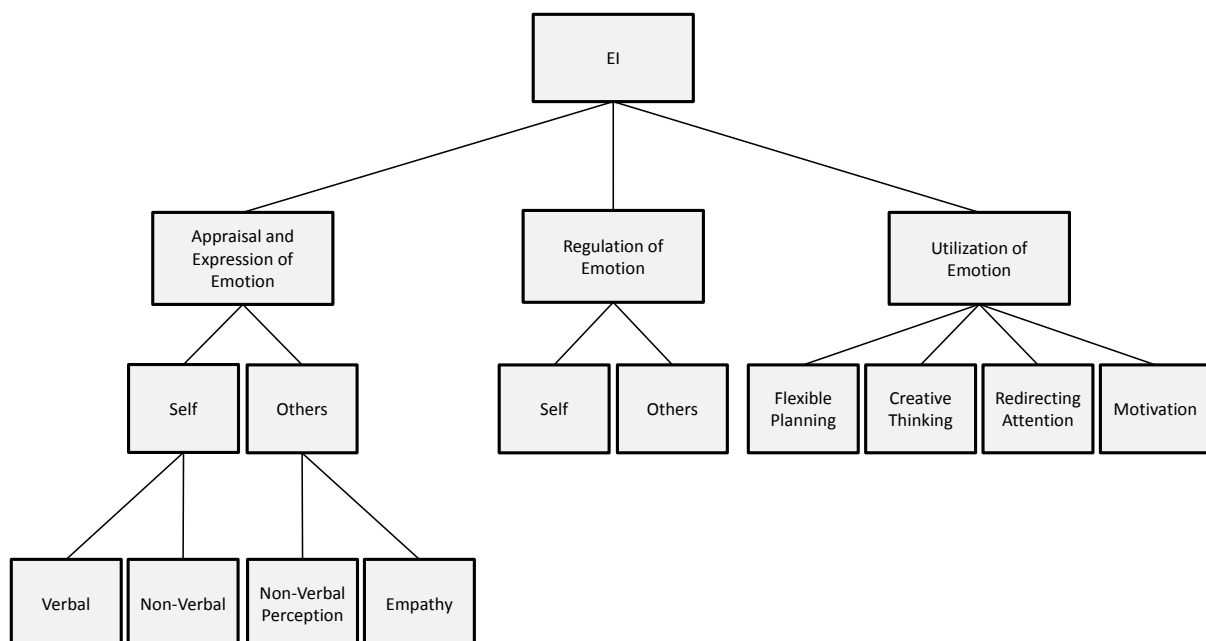


Abbildung 2.1. Das 1990-Modell der EI von Salovey und Mayer (1990).

### 2.1.2 Das Four-Branch Modell der EI

Angesichts der Schwächen des ersten Modells und auf Grundlage zwischenzeitlich neu veröffentlichter Studien, überarbeiteten die Autoren das 1990-Modell in den darauf folgenden Jahren (Mayer et al., 2000). Als Resultat ihrer Bemühungen stellten Mayer und Salovey im Jahre 1997 das sogenannte Four-Branch Modell der EI vor. Im Four-Branch Modell wird EI als ein Set von vier breiten Fähigkeitsbereichen aufgefasst, die von den Autoren des Modells als *Branches* bezeichnet wurden. Die vier Branches der EI umfassen a) die Fähigkeit Emotionen wahrnehmen und ausdrücken zu können (*Perception*), b) Emotionen generieren zu können, um das Denken zu erleichtern (*Facilitation*), c) Emotionen zu verstehen (*Understanding*) und d) Emotionen effektiv regulieren zu können (*Management*) (Mayer & Salovey, 1997; Mayer et al., 1999). Während jeweils mit den Branches Perception, Facilitation und Management ein Pendant der drei, im 1990-Modell postulierten Fähigkeiten zu finden ist, ist der Branch EU als neue Fähigkeit in diesem Modell hinzugekommen.

In Abbildung 2.2 wird das Four-Branch Modell illustriert. Die Autoren nahmen an, dass die vier Fähigkeiten eine schwache Hierarchie bilden, von basalen Fähigkeiten, die auf den unteren Ebenen des Modells dargestellt werden (z. B. Perception), bis hin zu komplexen Fähigkeiten weiter oben im Modell angeordnet sind (z. B. Management), die teilweise von basalen Fähigkeiten abhängen sollen (Mayer & Salovey, 1997; Mayer, Salovey, Caruso & Sitarenios, 2001).

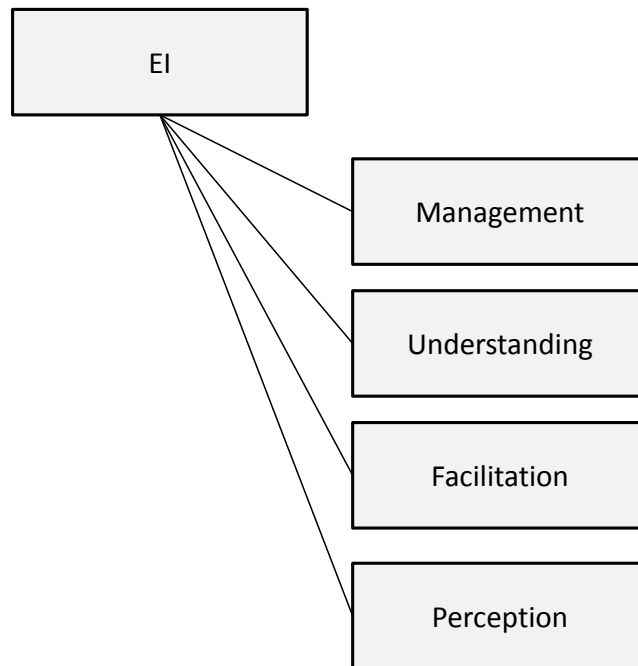


Abbildung 2.2. Das Four-Branch Modell der EI (Mayer & Salovey, 1997; Mayer et al., 1999).

Die vier Branches wurden von Mayer und Salovey (1997) als breite Fähigkeitsbereiche aufgefasst. Um diese Fähigkeiten näher zu beschreiben, schlugen die Autoren in ihrem Modell für jede der vier Branches eine Menge von vier Teilfähigkeiten vor, die repräsentativ für den jeweiligen Branch sein sollten. Ursprünglich wurde für diese Teilfähigkeiten zudem behauptet, dass sie einen Entwicklungsprozess innerhalb eines Branch reflektieren, von Fähigkeiten, die sich früh im Leben entwickeln, bis hin zu Fähigkeiten, die sich erst später im Leben ausbilden (Mayer & Salovey, 1997). Der angenommene Entwicklungsverlauf wurde bislang noch nicht untersucht. Zudem ist unklar, ob die repräsentativen Fähigkeiten trennbare Unterfähigkeiten innerhalb eines Bereichs der EI darstellen sollen. Später wurden die repräsentativen Fähigkeiten eher als Beispiele betrachtet, die zur Beschreibung der vier Fähigkeiten dienen sollen (siehe Mayer et al., 2000).

## *Perception*

Der Branch Perception repräsentiert die Fähigkeit, Emotionen korrekt identifizieren zu können (Mayer & Salovey, 1997). Mayer und Salovey schlugen als repräsentative Fähigkeiten für diesen Branch a) das korrekte Wahrnehmen der eigenen Emotionen, b) das korrekte Wahrnehmen der Emotionen Anderer, c) den akkuraten Ausdruck von Emotionen sowie d) die Fähigkeit, falsche und ehrliche Emotionsausdrücken voneinander unterscheiden zu können, vor. Die Fähigkeit Perception umfasst dabei die korrekte Wahrnehmung des Ausdrucks von Emotionen in Gesichtern, der Sprache, Körperhaltungen sowie vermittelt über Kunst oder Musik (Mayer et al. 2002; Mayer et al., 2008; Mayer & Salovey, 1997). Bei der Einführung des Four-Branch Modells wurde dieser Branch zunächst als *Perception and Expression* bezeichnet. In neueren Beschreibungen des Modells hat die Teilfähigkeit des akkuraten Ausdrucks von Emotionen an Bedeutung verloren und wird dort häufig nicht mehr genannt (siehe z. B. Mayer et al., 2008; Mayer, Salovey, Caruso & Sitarenios, 2003). Möglicherweise hängt das Verschwinden des Konzeptes aus dem Four-Branch Modell damit zusammen, dass die Fähigkeit zum akkuraten Ausdruck von Emotionen schwer standardisiert erfasst werden kann und die ersten Operationalisierungen des Modells durch die Autoren keine Untertests für diese Fähigkeit enthielten (siehe Mayer et al., 1999; Mayer et al., 2003). Außerhalb der Forschung zu EI gab es allerdings Bemühungen, diese Teilfähigkeit messbar zu machen. In der Diagnostic Analysis of Nonverbal Accuracy Scale (DANVA; Nowicki & Duke, 1994) ist ein Untertest enthalten, mit dem der akkurate Ausdruck von Emotionen weitestgehend standardisiert erfasst werden soll. In diesem Untertest besteht die Aufgabe der Testperson darin, eine Reihe vorgegebener Emotionen im Gesicht, der Gestik und der Stimme auszudrücken. Anschließend wird der Emotionsausdruck anhand von Videoaufzeichnungen von trainierten Auswertern hinsichtlich seiner Akkuratheit bewertet (Nowicki & Duke, 1994).

Die Fähigkeit sowie der Prozess der Emotionswahrnehmung stellen einen gut untersuchten Forschungsbereich dar, der bereits weit vor der Einführung des relativ jungen Konstruktes der EI begründet wurde (O`Sullivan & Ekman, 2004). Unglücklicherweise findet die Forschung zur EI und zur Emotionswahrnehmung häufig eher isoliert voneinander statt.

Für die Erfassung und theoretische Weiterentwicklung des Branch Perception könnte die Forschung zur Emotionswahrnehmung wichtige Hinweise liefern. In diesem Zusammenhang sind besonders die bisherigen Bemühungen interessant, a) den Ausdruck von Emotionen systematisch zu kodieren sowie b) Instrumente zur Erfassung von Emotionswahrnehmungsfähigkeit zu konstruieren. Für die Kodierung des Ausdrucks von Emotionen in Gesichtern haben Ekman und Friesen u. a. das Facial Action Coding System (FACS) entwickelt (Ekman & Oster, 1979). In diesem Kodierungssystem wird für verschiedene Emotionen das sichtbare Muster an Aktivierungen der Gesichtsmuskulatur beschrieben (Ekman & Oster, 1979; für einen Überblick siehe Matsumoto, Keltner, Shiota, O`Sullivan & Frank, 2008). Ein solches Kodierungssystem könnte auch eine Grundlage für das Scoring von Tests zur Erfassung von Perception bieten. Außerhalb der EI wurde das FACS bereits zu diesem Zweck verwendet. Im Japanese and Caucasian Brief Affect Recognition Test (JACBART; Matsumoto et al., 2000) werden den Testpersonen Videoaufnahmen von Gesichtern gezeigt, die hinsichtlich zweier ethnischer Gruppen sowie des Geschlechts ausbalanciert wurden. Die Emotionsausdrücke werden im JACBART für einen kurzen Zeitraum (d. h. 0.20 Sekunden) zwischen dem neutralen Gesichtsausdruck der gleichen Darstellerin bzw. Darstellers präsentiert (Matsumoto et al., 2000). Die Aufgabe der Testperson besteht darin, die gezeigten Emotionen zu identifizieren. Bei der Bewertung der Testantworten wird das FACS als Kriterium der Akkuratheit verwendet.

Die Forschung zur Emotionswahrnehmung ist allerdings nicht auf den Ausdruck von Emotionen in Gesichtern beschränkt. In einigen Ansätzen wurde dem Umstand Rechnung

getragen, dass Emotionen auch über andere Modalitäten wahrgenommen werden können (z. B. Bänziger, Grandjean & Scherer, 2009; Scherer & Scherer, 2011). Beispielsweise konnten akustische Parameter in Stimmen identifiziert werden, die bei der Wahrnehmung der Intensität und Qualität einer Emotion eine Rolle spielen (Banse & Scherer, 1996)

Mit dem Multimodal Emotion Recognition Test (MERT; Bänziger et al., 2009) wurde zudem ein Instrument zur Verfügung gestellt, das Emotionswahrnehmungsfähigkeit über mehrere Modalitäten erfassen soll. Dazu wurden Videoaufnahmen von professionellen Schauspielerinnen und Schauspielern erstellt, die vorgegebene Emotionen über das Gesicht, die Stimme und Körperhaltung ausdrücken. Den Getesteten werden die Emotionsausdrücke auf vier verschiedene Arten vorgelegt: als bewegtes Bild mit Ton, ausschließlich als Ton, ausschließlich als bewegtes Bild oder als unbewegtes Bild (Bänziger et al., 2009). Entsprechend dieser Art der Darstellung werden im MERT vier Scores der Akkuratheit gebildet.

Darüber hinaus gab es mit der Erstellung des Geneva Multimodal Emotion Portrayals Core Sets (GEMEP-CS; Bänziger, Mortillaro & Scherer, 2012) Bemühungen, einen umfassenden Korpus von Stimulus-Material für weitere Forschung zur Emotionswahrnehmungsfähigkeit zur Verfügung zu stellen, das hinsichtlich seiner Konstruktion mit dem Material des MERT vergleichbar ist. Dael, Mortillaro und Scherer (2012) haben versucht, auf Grundlage dieses Stimulus-Materials mit dem sogenannten Body Action and Posture (BAP) Systems Elemente von Körperbewegungen beim Emotionsausdruck zu identifizieren.

Die Entwicklung von Tests zur multimodalen Erfassung von Emotionswahrnehmungsfähigkeit könnte auch für die konzeptuelle Weiterentwicklung der Branch Perception eine Grundlage bieten. Eine Studie von Schlegel, Grandjean und Scherer

(2012) lieferte erste Hinweise darauf, dass Emotionswahrnehmungsfähigkeit nicht eindimensional ist, sondern modalitäts- und emotionsspezifisch sein könnte.

Einen weiteren Aspekt der Emotionswahrnehmungsforschung, der im Rahmen der EI bislang kaum betrachtet wurde, stellen die sogenannten Kontexteffekte der Emotionswahrnehmung dar. Im Rahmen dieser Forschung werden verschiedene Arten von Kontexten diskutiert, die einen Einfluss auf die Wahrnehmung haben können (siehe Barrett, Mesquita & Gendron, 2011). Eine Gruppe von Kontexteffekten bezieht sich auf Hinweise auf die Situation, in der die Emotion aufgetreten ist (Barrett et al., 2011). Möglicherweise könnte für diese Art von Kontexteffekten auch EU eine Rolle spielen, wenn die Situationen Hinweise auf die Gründe der Emotionen liefern.

Trotz dieser interessanten Entwicklungen im Bereich der Emotionswahrnehmungsfähigkeit, wurden nur selten Instrumente, die außerhalb der EI entwickelt wurden, und Operationalisierungen der Branch Perception miteinander in Verbindung gesetzt. In einer Studie von Roberts, Schulze, O'Brien, MacCann, Reid und Maul (2006) wiesen Operationalisierungen für Perception nicht die erwarteten Zusammenhänge mit Maßen für Emotionswahrnehmungsfähigkeit außerhalb der EI-Forschung auf. Diese fehlende Konvergenz ließ Zweifel an der Validität der bisherigen Operationalisierungen von Perception im Rahmen der EI aufkommen (Roberts et al., 2006). Das Ergebnis legt zudem nahe, dass das Konzept und die Erfassung von Perception zukünftig mehr im Rahmen der Emotionsforschung betrachtet werden sollte.

### ***Facilitation***

Der Branch Facilitation wurde als Fähigkeit definiert, Emotionen nutzen zu können, um kognitive Prozesse zu unterstützen (Mayer et al., 2008). Für Facilitation wurde ein Set von teilweise recht heterogenen repräsentativen Fähigkeiten vorgeschlagen. Ein erster Aspekt von Facilitation ist die Fähigkeit, Emotionen nutzen zu können, um die Aufmerksamkeit auf

wichtige Informationen zu lenken (Mayer & Salovey, 1997). Beispielsweise könnte eine Person die Emotion Besorgnis nutzen, um die Aufmerksamkeit auf Aktivitäten zu lenken (z. B. Lernen), mit denen ein unerwünschtes Ereignis (z. B. Nichtbestehen einer Prüfung) abgewendet werden kann (siehe Mayer & Salovey, 1997). Als zweite repräsentative Fähigkeit wurde vorgeschlagen, dass Personen Emotionen generieren können, um Emotionen verstehen zu können (Mayer & Salovey, 1997). Anhand der generierten Emotionen lassen sich die Gefühle einer anderen Person nachempfinden oder eigene Emotionen in zukünftigen Ereignissen vorhersagen (Mayer & Salovey, 1997). Ein dritter Aspekt von Facilitation bezieht sich darauf, durch Veränderungen der Stimmungen unterschiedliche Perspektiven einnehmen zu können und für Entscheidungen nutzen zu können. Als vierte repräsentative Fähigkeit wurde das Verstehen von differentiellen Effekten der Stimmung auf verschiedene kognitive Aufgaben vorgeschlagen.

Die Grundlage für den Vorschlag dieser Fähigkeit lässt sich auf eine Reihe von Studien außerhalb der EI zurückführen, in denen der Einfluss von Emotionen und Stimmungen auf kognitive Prozesse untersucht wurden (für eine Übersicht siehe Blanchette & Richards, 2010). Beispielsweise wurde der differentielle Effekt von positiven und negativen Stimmungen auf induktives und deduktives Reasoning untersucht sowie der Effekt von Stimmungen auf Prozesse des Urteilens und Entscheidens (Blanchette & Richards, 2010; Palfai & Salovey, 1993). Diese Effekte wurden allerdings selten im Zusammenhang mit einer Fähigkeit betrachtet. Palfai und Salovey (1993) schlugen vor, dass Personen Wissen über Effekte von Emotionen auf Kognitionen nutzen könnten, um ihre Leistung zu steigern, indem sie hilfreiche Emotionen herbeiführen. Diese angenommene Fähigkeit wurde bislang kaum untersucht.

Das Generieren von Emotionen, um Emotionen zu verstehen (die zweite repräsentative Fähigkeit) lässt sich hingegen mit einem anderen Forschungsbereich in Verbindung setzen.



Im Rahmen der Empathieforschung und der sogenannten Simulationstheorien (z. B. Batson, 2009) wird angenommen, dass Individuen die Perspektive anderer Personen einnehmen können, indem sie deren Gefühle simulieren. Dieser Aspekt der Empathie wird auch als affektive Perspektivübernahme bezeichnet (z. B. Oswald, 1996).

Aufgrund dieser verschiedenartigen vorgeschlagenen Fähigkeiten lässt sich der Kern von Facilitation nur schwer fassen. Gemeinsam ist den vorgeschlagenen Fähigkeiten, dass Emotionen eine unterstützende Rolle zugeschrieben wird, was Facilitation von den anderen Branches des Modells unterscheidet (Mayer et al., 2001). Im Hinblick auf die nähere Beschreibung durch die repräsentativen Fähigkeiten erscheint es zudem schwierig, Facilitation konzeptuell von anderen Branches eindeutig trennen zu können. Die Fähigkeit, Emotionen zu simulieren, um sich selbst oder andere zu verstehen (Mayer & Salovey, 1997) erscheint untrennbar von der Fähigkeit EU. Zudem erscheint es schwierig, eine Fähigkeit, die das willentliche Generieren von Emotionen beinhaltet, unabhängig von Management zu betrachten.

### ***Understanding***

Der dritte Fähigkeitsbereich im Four-Branch Modell ist EU. EU umfasst in diesem Modell die Fähigkeit, Emotionen verstehen zu können sowie Wissen über Emotionen (Mayer & Salovey, 1997). Nach Mayer und Salovey (1997) umfasst EU die Fähigkeiten, a) Emotionen korrekt benennen zu können und Ähnlichkeiten zwischen Emotionen zu erkennen, b) Verbindungen zwischen Situationen und der Auslösung von Emotionen verstehen zu können, c) Wissen über die Komplexität von Emotionen sowie d) Übergänge von einer Emotion zu einer anderen Emotion verstehen zu können.

Die erste repräsentative Fähigkeit ermöglicht es einem Individuum, zwischen Emotionsbegriffen diskriminieren zu können. Eine Person mit hoher Ausprägung auf EU könnte beispielsweise verstehen, dass den beiden Emotionsbegriffen Irritation und Wut

gemeinsam ist, dass sie der Emotion Ärger ähneln, sich dabei allerdings hinsichtlich ihrer Intensität voneinander unterscheiden.

Die zweite Fähigkeit bezieht sich auf das Verstehen von Verbindungen zwischen Situationen und Emotionen (Mayer & Salovey, 1997). Die Fähigkeit ermöglicht es Personen mit hoher EU-Fähigkeit akkurat vorherzusagen, welche Emotionen wahrscheinlich oder unwahrscheinlich in einer gegebenen Situation auftreten. Beispielsweise könnte eine Person verstehen, dass Verlust und Trauer miteinander einhergehen (Mayer & Salovey, 1997).

Die dritte Fähigkeit bezieht sich auf das Wissen über die Komplexität von Emotionen. Eine komplexe Emotion ist eine Emotion, die mehrere Emotionen beinhaltet. EU könnte beispielsweise dazu befähigen, zu verstehen, dass der Emotionsbegriff Erschrecken sowohl Furcht als auch Überraschung beinhaltet (Mayer & Salovey, 1997).

Die vierte repräsentative Fähigkeit für EU ist die Fähigkeit zum Schlussfolgern über Ketten von Emotionen (Mayer & Salovey, 1997). Ein Individuum mit hohem EU ist fähig, wahrscheinliche Übergänge von einer Emotion zu einer anderen Emotion sowie wahrscheinlich Veränderungen der Intensität von Emotionen über die Zeit vorherzusagen (Mayer et al., 2001).

Der Begriff EU wird auch außerhalb der Domäne EI für emotionsbezogene Fähigkeiten verwendet, vor allem in entwicklungspsychologischen Studien (siehe Harris, 2008). Allerdings ist der Begriff EU hier meist breiter als im Four-Branch Modell gefasst und beinhaltet, neben dem Verstehen von Ursachen von Emotionen, auch die Fähigkeiten, Emotionen wahrnehmen und regulieren zu können (siehe Pons, Harris & de Rosnay, 2004). Trotz der Verwendung des gleichen Labels, können somit sehr unterschiedliche Fähigkeiten innerhalb und außerhalb der Forschung zu EI unter EU verstanden werden. Ein weiterer Begriff, der häufig in entwicklungspsychologischen Studien verwendet wird und teilweise

Überschneidungen mit EU hat, ist Emotion Knowledge (Izard, Trentacosta, King, Morgan & Diaz, 2007). Emotion Knowledge umfasst das Wahrnehmen und korrekte Benennen von Emotionen, das Verstehen von komplexen Emotionen und das Verstehen von Gründen für Emotionen (Izard et al., 2007). Unter Emotion Knowledge werden also Fähigkeiten gefasst, die im Rahmen der EI durch Perception und EU repräsentiert sind (Izard et al., 2007).

EU, wie es im Rahmen der EI definiert wurde, hat deutliche Überschneidungen mit Konzepten aus der Emotionsforschung, die wiederum anders benannt wurden. Hier sind vor allem die gut untersuchten Konzepte der Empathie und des Affectiven Forecastings zu nennen. Aufgrund der Bedeutung für EU, werden sie in einem späteren Kapitel näher beleuchtet.

Betrachtet man den Branch EU fällt auf, dass die genannten Aspekte teilweise recht unterschiedlich sind. Die Autoren des Four-Branch Modells fassten die repräsentativen Fähigkeiten von EU später unter zwei inhaltlichen Aspekten zusammen (siehe Mayer et al., 2000, 2001): Zum einen umfasst EU das Wissen über die Bedeutung von Emotionsbegriffen, das die Teilfähigkeiten des Benennens von Emotionen und das Verstehen von gemischten Emotionen beinhaltet. Hierbei handelt es sich um die linguistische Bedeutung von Emotionsbegriffen (Mayer et al., 2001). Zum anderen enthält EU das Verstehen des Entstehens und der Veränderungen von Emotionen. Die Konzeptualisierung von EU im Four-Branch Modell ist somit nur teilweise mit der Definition von EU vereinbar, wie sie in dieser Arbeit verwendet wird. Während der zweite Teilaspekt (das Verstehen des Entstehens und der Veränderungen von Emotionen) den Kern der Definition von EU in dieser Arbeit trifft, gilt dies für den ersten Aspekt (das linguistische Wissen) nur sehr eingeschränkt. Allein das Wissen über bestimmte Teilaspekte einer Emotion, die sich auf die Auslösung und den Verlauf von Emotionen beziehen, können nach der hier verwendeten Definition ein Teil von EU sein. Beispielsweise kann das Wissen über den Emotionsbegriffs „Erleichterung“ dessen

Implikation umfassen, dass eine unerwünschte Situation nicht eingetroffen ist bzw. eine negative Emotion zurückgegangen ist. Das Wissen über andere Aspekte der Bedeutung eines Emotionsbegriffs, wie bspw. die Komplexität oder Intensität einer Emotion, wird in der hier genannten Definition nicht unter EU gefasst. Insgesamt erscheinen die beiden von Mayer et al. (2000) vorgeschlagenen inhaltlichen Aspekte von EU so unterschiedlich, dass es einer genaueren Untersuchung bedarf, ob diese beiden Aspekte wirklich Teile einer einzigen Fähigkeit, d. h. EU, sind.

### ***Management***

Der vierte Branch, Management, stellt die Fähigkeit dar, Emotionen regulieren zu können. Management umfasst die Fähigkeit a) Gefühlen offen gegenüber zu sein, b) Emotionen zu einem angemessenen Zeitpunkt zu erleben, c) Emotionen bewusst zu reflektieren sowie d) die eigenen und die Emotionen Anderer zu regulieren (Mayer & Salovey, 1997).

Unter der Fähigkeit, Gefühlen offen gegenüber zu stehen, verstanden Mayer und Salovey (1997) das Zuwenden von Aufmerksamkeit zu Emotionen unabhängig davon, ob sie angenehm oder unangenehm sind. Mayer und Salovey (1997, S. 14) bezeichneten dies als „openness to feelings“. Insgesamt ist die Beschreibung dieser Fähigkeit im Four-Branch Modell sehr vage. Eine Bezeichnung als Offenheit für Gefühle weist zudem eine deutliche Nähe zu der Domäne der Persönlichkeit oder Motivation auf, die im Rahmen eines Fähigkeitsmodells unangemessen ist.

Die zweite repräsentative Fähigkeit ermöglicht es Personen, die Angemessenheit von Emotionen einschätzen zu können sowie eine Emotion zu einem angemessenen Zeitpunkt erleben oder sich von ihr lösen zu können (Mayer & Salovey, 1997). Die Angemessenheit der Emotion kann durch soziale Normen vorgegeben sein oder durch Erwartungen über den Nutzen von Emotionen (Mayer & Salovey, 1997). Beispielsweise könnten in bestimmten

Situationen eine Reaktion mit Ärger unangemessen sein und für das Individuum zu negativen Konsequenzen führen.

Unter der dritten Fähigkeit verstanden die Autoren, dass ein Individuum über seine Emotionen bewusst reflektieren kann. In diesem Prozess werden die Emotionen dahingegen bewertet, wie typisch, einflussreich oder vernünftig sie in einer bestimmten Situation sind (Mayer & Salovey, 1997).

Die vierte repräsentative Fähigkeit erlaubt es Personen, die eigenen Emotionen und die Emotionen Anderer zu regulieren. Mayer und Salovey (1997) führten in diesem Kontext das willentliche Mindern von negativen Emotionen sowie das Erhöhen und Mindern von positiven Emotionen auf. Côté, DeCelles, McCarthy, Van Kleef und Hideg (2011) schlugen hingegen vor, dass dazu auch die Erhöhung negativer Emotionen zählen könnte und Emotionsregulation eine negative Seite haben kann. Die Autorinnen und Autoren konnten zeigen, dass Emotionsregulationsfähigkeit, abhängig von den Persönlichkeitseigenschaften einer Person, auch interpersonal-deviantes Verhalten erleichtern kann (Côté et al., 2011).

Außerhalb der EI weist der Branch Management deutliche Überschneidungen mit dem Forschungsbereich der Emotionsregulation auf (für einen Überblick siehe Gross, 2008). In dieser Domäne steht der Aspekt von Management im Vordergrund, der sich auf die Regulierung eigener Emotionen bezieht. Theorien der Emotionsregulation beschäftigen sich zum einen mit motivationalen Aspekten, die hinter der Emotionsregulation stehen und zum anderen mit der Frage, wie Emotionen reguliert werden können (Niedenthal, Krauth-Gruber & Ric, 2006). Letzteres erscheint auch für EI relevant zu sein. Ein prominentes Modell der Emotionsregulation ist das Prozess-Modell der Emotionsregulation von Gross (1998). In diesem Modell wird angenommen, dass Personen ihre Emotionen mit Hilfe bestimmter Strategien regulieren, die zu verschiedenen Zeitpunkten des Emotionsauslösungsprozesses eingesetzt werden. Gross (1998) unterschied zwei Arten von Strategien, entsprechend dem

Zeitpunkt, wann sie zum Einsatz kommen: Strategien können entweder angewendet werden bevor die Emotion vollständig ausgelöst wurde (*Antecedent-Focused*) oder nachdem die Emotion ausgelöst wurde (*Response-Focused*). Zur Antecedent-Focused Emotionsregulation werden Strategien gezählt, wie die Auswahl der Situation (*Situation Selection*), die Veränderung der Situation (*Situation Modification*), die Lenkung von Aufmerksamkeit (*Attention Deployment*) oder die kognitive Bewertung der Situation (*Reappraisal*) (siehe Gross, 1998). Zur Response-Focused Emotionsregulation gehört die Unterdrückung des emotionalen Ausdrucks, die Unterdrückung von Gedanken und das soziale Teilen von Emotionen mit anderen Personen (Gross, 1998). Neben der Beschreibung von Emotionsregulations-Strategien, gab es auch große Bemühungen, die Effektivität dieser Strategien zu untersuchen (siehe Webb, Miles & Sheeran, 2012). In einer Meta-Analyse von Webb et al. (2012) erwies sich vor allem Reappraisal als effektive Strategie über verschiedene Studien hinweg.

Das Modell von Gross (1998) bietet zwar eine Beschreibung, mit welchen Strategien Personen ihre Emotionen regulieren - allerdings stellt sich die Frage, wie Personen diese Strategien in einer konkreten Situation auswählen (Gross, 2008). Die Wahl bestimmter Emotionsregulations-Strategien wird oft als gewohnheitsmäßiges Verhalten im Sinne eines Stils betrachtet und wird in diesem Zusammenhang über Selbstberichte erfasst (z. B. Gross & John, 2003). Gross (2008) ging allerdings davon aus, dass Personen ihre Emotionsregulations-Strategien auch auf Grund von Überzeugungen über deren Effektivität wählen können. In diesem Zusammenhang wurde Emotionsregulation auch im Sinne einer Fähigkeit diskutiert, die es ermöglicht, dass Personen effektive Strategien wählen. Mit der Intention, diese Fähigkeit messbar zu machen, wurde das Emotion Regulation Profile Questionnaire (ERP-Q; Mikolajczak, Nelis, Hansenne & Quoidbach, 2008) entwickelt. Im ERP-Q werden den Testpersonen 12 Situationsbeschreibungen mit je sechs möglichen Reaktionen präsentiert, die jeweils bestimmte Emotionsregulationsstrategien widerspiegeln. Die Aufgabe der Testperson

ist es, je zwei Reaktionen zu markieren, die sie am wahrscheinlichsten zeigen würden sowie zwei, die sie am unwahrscheinlichsten wählen würden. Die Strategien wurden zuvor anhand von Erkenntnissen aus der Literatur über die Effektivität von Emotionsregulations-Strategien als entweder funktional oder dysfunktional klassifiziert (Mikolajczak et al., 2008). Als Indikator für die Fähigkeit wird gewertet, wenn eine Person einer funktionalen Strategie zustimmt oder eine dysfunktionale Strategie ablehnt. Diese Klassifikation von Strategien hinsichtlich ihrer Effektivität ist allerdings nicht unproblematisch. Bonanno und Burton (2013) wiesen darauf hin, dass die Effektivität von Emotionsregulations-Strategien nicht generell gilt, sondern von Merkmalen der Situation (wie bspw. dem Zeitpunkt während des Emotionsauslösungsprozesses) abhängig ist.

Bei der Frage, wie Personen, abhängig von der Situation, effektive Emotionsregulations-Strategien wählen, könnte EI eine Rolle spielen (siehe Wranik, Feldman Barrett & Salovey, 2007). Wranik et al. (2007) schlugen vor, dass EU eine Grundlage für die Wahl effektiver Emotionsregulation liefern kann, da EU Wissen über die Emotionsauslösung bereithält. Côté et al. (2011) bezeichneten die Fähigkeit, die effektivste Emotionsregulations-Strategie in einer bestimmten Situation zu wählen als *Emotion-Regulation Knowledge*. Im Rahmen dieses Erklärungsansatzes muss zukünftig allerdings geklärt werden, was der Branch Management bzw. die Emotionsregulationsfähigkeit über EU hinaus erklärt. Zudem wurde angezweifelt, dass Wissen über die Effektivität von Emotionsregulations-Strategien allein ausreicht, um eine Emotion in emotionsgeladenen Situationen tatsächlich regulieren zu können (Nelis, Quoidbach, Mikolajczak & Hansenne, 2009; siehe auch Freudenthaler & Neubauer, 2005).

### **2.1.3 Struktur des Four-Branch Modells**

Mayer und Salovey (1997; siehe auch Mayer et al., 2000) fassten die vier Branches als breite Fähigkeitsbereiche auf, die zusammen eine Hierarchie bilden (Mayer et al., 2001). Über die

konkrete faktorielle Struktur der EI wurden verschiedene Vorschläge gemacht (siehe Wilhelm, 2005). Anhand der bisherigen Operationalisierungen des Four-Branch Modells, dem MEIS und MSCEIT, wurde untersucht, ob es einen General-Faktor der EI gibt und ob Evidenz für die vier Branches gefunden werden kann. Die Ergebnisse dieser Studien führten zu unterschiedlichen Schlussfolgerungen darüber, ob das angenommene Modell mit vier Faktoren eine adäquate Beschreibung der EI ist und führten zu Modifikationen des Modells. Die angenommene Hierarchie der Branches wurde hingegen kaum untersucht.

### ***2.1.3.1 Faktorielle Struktur***

Mayer et al. (1999) untersuchten die Struktur des Four-Branch Modells auf Grundlage von Daten für den MEIS. Im Rahmen dieser Untersuchung fanden die Autoren nur schwache Evidenz für den Faktor Facilitation. Die Ladungsmuster von Untertests für den Faktor Facilitation entsprachen in einer exploratorischen Faktorenanalyse nicht den Erwartungen, die aus dem Four-Branch Modell abgeleitet werden können (Mayer et al., 1999). In einer anschließenden konfirmatorischen Faktorenanalyse waren die Faktoren Understanding und Facilitation so hoch miteinander korreliert, dass sie kaum von voneinander trennbar erschienen, was die Autoren als einen Hinweis für ein Modell der EI mit drei Faktoren werteten. Auch in einer Studie von Roberts et al. (2001) entsprachen die Ladungsmuster der Untertests des MEIS nicht den Erwartungen. Ein anschließender Modellvergleich wies allerdings darauf hin, dass ein Modell mit vier Faktoren eine signifikant bessere Passung aufwies, als ein Modell mit nur drei Faktoren (Roberts et al., 2001). Ciarrochi, Chan und Caputi (2000) hingegen schlugen, ebenfalls anhand von Daten für den MEIS, eine weitere Alternative für das Four-Branch Modell vor. Aus Sicht der Autorinnen und Autoren deuteten die Daten auf ein 2-faktorielles Modell hin, dessen erster Faktor als Perception und dessen zweiter Faktor als Understanding und Management interpretiert werden kann. Insgesamt kann somit festgehalten werden, dass die Evidenz für die Struktur des Four-Branch Modells,



beurteilt anhand des MEIS, als eher schwach bezeichnet werden kann, wobei vor allem Unklarheit bezüglich des Branch Facilitation besteht.

Basierend auf Daten für den MSCEIT und in Bezugnahme auf das von Ciarrochi et al. (2000) vorgeschlagene 2-faktorielle Modell, schlugen Mayer et al. (2002) eine Erweiterung des Four-Branch Modells vor. In dieser Erweiterung wurde eine weitere Ebene in das Four-Branch Modell eingeführt, mit zwei Faktoren, die den vier Branches übergeordnet sind. Diese beiden Faktoren wurden als *Experiential* und *Strategic EI* bezeichnet. Während *Experiential EI* den beiden Faktoren *Perception* und *Facilitation* übergeordnet ist, umfasst *Strategic EI* die Branches *Understanding* und *Management*. Die Interpretation der beiden neuen Faktoren geht dabei nicht weit über eine Beschreibung der jeweils einzelnen Branches hinaus. Mayer et al. (2002) konzeptualisierten *Experiential EI* als “ability to perceive emotional information, to relate it to other sensations [...], and to use it to facilitate thought” (S. 17). *Strategic EI* wurde definiert als „ability to understand emotional information and use it strategically for planning and self-management“ (Mayer et al., 2002, S. 17). Mayer et al. (2003) fanden auf Grundlage des MSCEIT eine sehr gute Passung für das Four-Branch Modell sowie erste Hinweise für die beiden Faktoren *Experiential* und *Strategic EI*.

Die Ergebnisse von Mayer et al. (2002, 2003) konnten in anderen Studien nur teilweise bestätigt werden (Palmer, Gignac, Manocha & Stough, 2005; Rossen, Kranzler & Algina, 2008). Palmer et al. (2005) und Rossen et al. (2008) verglichen mehrere theoretisch mögliche Modelle der EI miteinander. Beide Studien kamen zu dem Ergebnis, dass die Daten nicht allein mit einem General-Faktor der EI erklärt werden können und dass die Vier-Faktoren-Lösung nicht die beste Erklärung der Daten ist, da die Evidenz für den Faktor *Facilitation* nur schwach war. In beiden Studien ergab ein Modell mit einem General-EI-Faktor und zusätzlichen drei Faktoren *Perception*, *EU* und *Management* (wobei die Korrelation zwischen *EU* und *Management* zugelassen ist) die beste Passung auf die Daten.

Fan, Jackson, Yang, Tang und Zhang (2010) konnten die Ergebnisse größtenteils in einem meta-analytischen Strukturgleichungsmodell (siehe Cheung & Chang, 2005) über 19 Studien bestätigt. Auch hier wies ein einfaktorielles Modell der EI ungenügenden Fit auf. In einem Modell mit vier Faktoren waren Perception und Facilitation mit .90 miteinander korreliert. Auch Fan et al. (2010) schlugen daher ein Modell mit den drei Branches Perception, Understanding und Management vor.

Insgesamt deuten sowohl die Ergebnisse zum MEIS als auch zum MSCEIT darauf hin, dass die Evidenz für den theoretisch kaum belegten Faktor Facilitation bisher unzureichend ist. Diese Schlussfolgerungen beruhen allerdings bislang nur auf dem MEIS/MSCEIT-Messansatz und sind damit auch abhängig von der Validität dieses Ansatzes.

### ***2.1.3.2 Hierarchie-Annahme***

Die angenommene Hierarchie zwischen den Branches der EI wurde zum Zeitpunkt dieser Arbeit noch nicht belegt und inhaltliche Überlegungen lassen zudem Zweifel daran aufkommen, dass einige der Fähigkeiten als Voraussetzungen für höhere Branches betrachtet werden können. Zwei theoretische Beispiele sollen dies illustrieren:

Zwar ist es möglich, dass eine Emotion zunächst wahrgenommen und aufgrund der Wahrnehmung anschließend reguliert wird. Andererseits könnte eine Emotion aber auch bereits vor ihrem Auftreten reguliert werden (siehe Gross, 2008). Im Rahmen von frühen Regulations-Strategien, wie dem situation selection, wäre die Wahrnehmung der Emotion nicht notwendig und die Vorhersage, dass die Emotion auftreten wird, ausreichend.

Ebenso erscheint es möglich, dass Emotionen verstanden werden, ohne dass das Wahrnehmen von Emotionen involviert ist. Dies ist möglich, wenn Personen direkt über ihre emotionalen Reaktionen sowie deren Gründe der Auslösung mit Anderen sprechen. Dass Personen tatsächlich häufig ihre emotionalen Reaktionen mit anderen Personen teilen, konnte

in mehreren Studien gezeigt werden (Rimé, Mesquita, Philippot & Boca, 1991). Aufgrund der gemessenen Häufigkeit, wird das soziale Teilen von Emotionen von einigen sogar als ein charakteristischer Teil einer emotionalen Reaktion betrachtet (Rimé et al., 1991). Einige Studien weisen darauf hin, dass dieses Teilen von Emotionen direkte Auswirkungen auf den Prozess des Verstehens von Emotionen haben kann. Die Häufigkeit des Sprechens über kausale Ursachen von Emotionen mit Eltern oder Peers ist mit der späterer Fähigkeit, Emotionen zu verstehen, positiv assoziiert (Dunn, Brown, Slomkowski, Tesla & Youngblade, 1991; Hughes & Dunn, 1998), so dass zumindest ein Teil des Emotionsverständnisses auf diese Art aufgebaut werden könnte. Somit kann die Wahrnehmung von Emotionen dem Verstehen von Emotionen vorangehen: es wäre aber auch denkbar, dass Emotionen verstanden werden können, ohne diese wahrgenommen zu haben.

In einer Meta-Analyse von Joseph und Newman (2010) zum Zusammenhang von EI und beruflicher Leistung wurde u. a. auch die Annahme der Hierarchie der vier Branches diskutiert. Joseph und Newman (2010) führten in einem Pfadmodell die berufliche Leistung auf die vier Branches der EI, Gewissenhaftigkeit, emotionaler Stabilität sowie kognitive Fähigkeiten zurück. Innerhalb dieses Modells wurde eine Regression von Management auf Understanding geschätzt sowie eine Regression von Understanding auf Perception. Joseph und Newman (2010) konnten zum einen erfolgreich inkrementelle Validitätsevidenz für EI bei der Vorhersage von beruflicher Leistung finden. Die Autoren zogen darüber hinaus aber auch die Schlussfolgerung, dass ein Beleg für die Hierarchie des Four-Branch Modells gefunden wurde. Dieser Schluss sollte allerdings mit Vorsicht betrachtet werden. Zum einen weisen die, in der Meta-Analyse aggregierten Primärstudien keine Designs auf, die tatsächlich eine kausale Schlussfolgerung erlauben, worauf die Autoren selbst einschränkend hinweisen (siehe Joseph & Newman, 2010). Zum anderen ist es denkbar, dass es eine Reihe äquivalenter Modelle gibt, die die Daten gleich gut erklären können, obschon die Autoren berichten, auch andere Modelle betrachtet zu haben (Joseph & Newman, 2010). Aufgrund der theoretischen

Überlegungen und der fehlenden empirischen Evidenz, sollte die häufig zitierte Annahme einer Hierarchie zukünftig genauer geprüft werden.

## **2.3 Abgrenzung zu Mixed Models und Trait EI**

Seit der Begriff EI in die wissenschaftliche Literatur eingeführt wurde, wurden neben dem Four-Branch Modell eine Vielzahl von weiteren Modellen zur Beschreibung der EI vorgeschlagen. Diese Modelle können als höchst heterogen bezeichnet werden, bezüglich der Annahmen darüber, wie EI definiert werden kann und welche Konzepte darunter gefasst werden können. Ein wesentliches Unterscheidungsmerkmal von Modellen der EI bezieht sich auf eine Art der Klassifikation von Konstrukten und deren Operationalisierungen, die zuerst von Cronbach eingeführt wurde (z. B. Cronbach, 1990).

### **2.3.1 Maximales versus Typisches Verhalten**

Cronbach (z. B. Cronbach, 1990) schlug die Unterscheidung zwischen dem sogenannten *maximalen* und *typischen Verhalten* vor. Nach dieser Vorstellung wird maximales Verhalten durch Fähigkeitskonstrukte bestimmt. Operationalisierungen maximalen Verhaltens sind dadurch charakterisiert, dass bei der Testung von der Testperson die bestmögliche Leistung verlangt wird (Cronbach, 1990). Typisches Verhalten hingegen umfasst das charakteristische Handeln und Empfinden einer Person in wiederkehrenden Situationen oder Klassen von Situationen, worunter Verhaltensstile, Persönlichkeitseigenschaften und Interessen gefasst werden können (Cronbach, 1990). Typisches Verhalten kann in vielen Fällen sinnvoll durch Selbstberichtsverfahren operationalisiert werden, da häufig davon ausgegangen werden kann, dass Personen einen guten Einblick in das eigene Erleben und Verhalten haben (Cronbach, 1990).

Im Rahmen des Four-Branch Modells ist EI als ein Set mentaler Fähigkeiten konzeptualisiert worden (Mayer et al., 1999; Mayer et al., 2008). In Bezug auf die

Klassifikation von Cronbach, beeinflusst EI somit das maximale Verhalten eines Individuums. Dementsprechend gab es Bemühungen, leistungsorientierte Tests der EI zu entwickeln. Daneben gibt es aber auch eine große Anzahl von Modellen, in denen EI als eine Mischung aus diversen emotionsbezogenen Persönlichkeitseigenschaften, Interessen, motivationalen Konzepten sowie Selbsteinschätzungen von Fähigkeiten beschrieben wird. Hier wird EI hauptsächlich oder ausschließlich als Form des typischen Verhaltens aufgefasst und über Selbstberichte operationalisiert. Beispiele für Konstrukte, die in diesem Kontext vorgeschlagen wurden, sind Impulsivität, Optimismus oder Happiness (siehe Petrides & Furnham, 2003). Dazu gehören auch Selbsteinschätzungen von Fähigkeiten. Beispielsweise wird Emotionsregulation hier mit Items erfasst wie, „When someone offends me, I'm usually able to remain calm“ (Petrides & Furnham, 2003).

Ein wichtiger Schritt in der Forschung zur EI war die Feststellung der Heterogenität von Annahmen über EI. Wie in der Einleitung erwähnt, wurden zwei Arten der Klassifikation von Modellen der EI vorgeschlagen: a) Ability Models versus Mixed Models (Mayer et al., 2000) und b) Ability-EI versus Trait-EI (Petrides & Furnham, 2000).

### **2.3.2 Mixed Models**

Die erste Art der Klassifikation wurde von Mayer et al. (2000) eingeführt. Die Autoren unterschieden die sogenannten Ability Models der EI von Mixed Models. Ability Models sind Modelle, in denen EI, wie im Four-Branch Modell, auf Fähigkeiten bzw. maximales Verhalten begrenzt ist. In den Mixed Models hingegen, wird EI als eine Mischung aus Konzepten aus der Fähigkeits- und Persönlichkeitsdomäne definiert und determiniert somit auch das typische Verhalten. Mayer et al. (2000) betonten, dass Mixed Models keine geeignete Vorstellung einer neuen Form von Intelligenz darstellen können. Bis heute stellt das Four-Branch Modell das einzige Ability Model der EI dar.

### 2.3.3 Trait EI

Eine zweite Art der Klassifikation von Konzeptionen der EI wurde von Petrides und Furnham (2000) vorgeschlagen. Im Rahmen dieses Vorschlags wurde die Existenz zweier unterschiedlicher Konstrukte der EI postuliert: Ability-EI und Trait-EI. Ability-EI wurde auch als *information-processing EI* bezeichnet und entspricht in dieser Klassifikation dem, was Mayer et al. (2000) unter Ability Models der EI fassten. Trait-EI hingegen soll eine weitere Art der EI darstellen, die das typische Verhalten einer Person über eine Vielzahl von Situationen hinweg bestimmt (Petrides & Furnham, 2000). Trait-EI wurde als *emotional self-efficacy* umschrieben und umfasst emotionsbezogene Selbstwahrnehmungen und Dispositionen (Petrides & Furnham, 2003). Petrides und Furnham (2003) argumentierten, dass der wesentliche Unterschied zwischen Ability- und Trait-EI nicht in der Theorie per se zu finden ist, sondern im Messansatz begründet liegt. Ability-EI sollte mit Maßen des maximalen Verhaltens, d. h. Leistungstests, erfasst werden (Petrides & Furnham, 2001). Unter der Verwendung von Selbstberichten wird nach Ansicht der Autoren Trait-EI gemessen. Dass die Theorie nicht ausschlaggebend sei, begründeten Petrides und Furnham (2001) damit, dass eine Operationalisierung des Four-Branch Modells als Leistungstest und als Selbstbericht zu unterschiedlichen Ergebnissen führen würde. Allerdings stellt sich hierbei die Frage, ob ein Selbstbericht überhaupt eine valide Möglichkeit darstellt, ein Fähigkeitskonstrukt zu operationalisieren.

Die Klassifikation von EI als Ability Models versus Mixed Models bzw. die Unterscheidung von Ability-EI und Trait-EI hat unterschiedliche Implikationen. Während Mayer et al. (2000) die Klasse der Mixed Models als ungeeignete Beschreibungen einer Form von Intelligenz betrachten, impliziert die Vorstellung einer Trait-EI, dass der Begriff EI auch adäquat für emotionsbezogene Persönlichkeitseigenschaften ist.

### **2.3.4 Kritik an Mixed Models und Trait-EI**

Schulze et al. (2007) kritisierten Mixed Models sowie das Konzept der Trait EI auf Grundlage von semantischen, theoretischen und empirischen Argumenten: Auf semantischer Ebene spricht gegen Mixed Models und Trait-EI, dass der Begriff *Intelligenz* nur für Konstrukte verwendet wird, wenn es sich dabei um kognitive Fähigkeiten handelt (Schulze et al., 2007). Für die Autoren kann damit eine Klassifikation, in denen EI einmal als Fähigkeit und einmal außerhalb von Fähigkeiten aufgefasst wird, nicht sinnvoll sein. Auf theoretischer Ebene führten die Autoren die Tatsache an, dass EI als Fähigkeit deutlich besser mit etablierten Intelligenzmodellen vereinbar sind. Zudem spricht empirisch gegen Mixed Models und Trait-EI, dass ihre Instrumente kaum mit Tests für Standard-Intelligenz oder Fähigkeitstests der EI korreliert sind (Schulze et al., 2007) und deutliche Überschneidungen mit Persönlichkeitseigenschaften aufweisen (siehe Roberts, Schulze & MacCann, 2008). Schulze et al. (2007) argumentierten, dass die Klassifikationen von Auffassungen der EI zu einer konzeptuellen Inkohärenz im Bereich EI beigetragen haben, die den Fortschritt in dieser Domäne erschwert. Trotz der Kritik an Mixed Models und Trait-EI, wurde darauf aufmerksam gemacht, dass die darunter gefassten Konstrukte wichtige Outcomes vorhersagen können und interessant für die Weiterentwicklung von Persönlichkeitstheorien sein könnten (Roberts et al., 2008; Schulze et al., 2007).

### **2.4 EI und Theorien der Intelligenz**

Da angenommen wird, dass EI eine neue Form der Intelligenz darstellt, sollte EI auch im Kontext bereits bestehender Theorien der Intelligenz betrachtet werden. Hierbei stellt sich die Frage, ob eine Fähigkeit der EI prinzipiell mit Modellen der Intelligenz vereinbar ist. Falls ja, ergibt sich zum anderen die Frage, in welcher Beziehung EI zu anderen Fähigkeiten in einem Modell der Intelligenz steht. Diese theoretischen Überlegungen erscheinen wichtig, zumal

von EI, im Sinne eines Kriteriums für eine neue Form der Intelligenz, erwartet wurde, dass sie positiv mit klassischer Intelligenz korreliert ist (Mayer et al., 1999).

### 2.4.1 Spearman's $g$

Eine der ersten und bis heute einflussreichsten Intelligenztheorien ist die Zwei-Faktorentheorie von Spearman. In Spearman's Zwei-Faktorentheorien wird angenommen, dass kognitive Leistungen von zwei Arten von Faktoren abhängen: von einem Generalfaktor  $g$ , der Leistungen in jeder mentalen Aufgabe beeinflusst und spezifischen Faktoren, die für jede Art von Aufgabe einzigartig sind (Spearman, 1920). Die Größe des Einflusses des gemeinsamen Faktors  $g$  kann zwischen verschiedenen Aufgaben variieren (Jensen, 1992). Diese Theorie wird teilweise auch als *Ein-Faktor-Theorie* bezeichnet, da angenommen wird, dass es nur einen gemeinsamen Faktor gibt (Carroll, 1993). Spearman's Theorie basiert auf der Beobachtung, dass Leistungstests aus unterschiedlichsten Domänen positiv miteinander korreliert sind (z. B. Hart & Spearman, 1912), was auch als Positive Manifold Phänomen bezeichnet wird (Jensen, 1992). Diese positiven Korrelationen werden in dieser Theorie auf den gemeinsamen Faktor  $g$  zurückgeführt.

Die Annahmen einer neuen Form von Intelligenz, die als EI bezeichnet werden kann, ist im Rahmen von Spearman's Theorie problematisch (Schulze, Roberts, Zeidner & Matthews, 2005). Die Vorstellung, dass jegliche Korrelation zwischen Leistungen durch einen einzigen gemeinsamen Faktor erklärt werden kann, lässt keinen Raum für ein Konzept der EI. Unter dieser Theorie würden Leistungen bei emotionsbezogenen Aufgaben sowie mögliche Korrelationen von EI-Tests mit anderen Leistungstests, auf  $g$  zurückgeführt werden. Auch wenn  $g$  als wichtiges Konzept gesehen wird und viel Evidenz für die Existenz eines gemeinsamen Faktors existiert (z. B. Jensen, 1992), haben andere die Wichtigkeit von Gruppenfaktoren betont (z. B. im Rahmen von Thurstones Primary Abilities) oder



hierarchische Modelle vorgeschlagen, bei denen der breiteste Faktor *g* auf höchster Ebene des Modells verortet wird (z. B. Carroll, 1993).

#### 2.4.2 Gf-Gc Theorie

Die *Gf-Gc Theorie* von Cattell stellt ebenfalls eine einflussreiche Intelligenztheorie dar (z. B. Cattell, 1963). Im Rahmen dieser Theorie wurde die Annahme einer einheitlichen Generalintelligenz, wie sie von Spearman vorgeschlagen wurde, angezweifelt (Cattell, 1963; Horn & Cattell, 1966). In der Gf-Gc Theorie werden zwei breite Faktoren der Intelligenz angenommen: fluide Intelligenz (Gf) und kristallisierte Intelligenz (Gc). Gf wurde als Fähigkeit zum Schlussfolgernden Denken bezeichnet und umfasst Leistungen, die die Anpassung auf neuartige Situationen erfordern (Cattell, 1963; Carroll, 1993). Gc hingegen wurde als das Ergebnis von Erfahrungen, Bildung und Akkulturation betrachtet (Horn & Cattell, 1966). Tests zur Erfassung von Gc erfordern das Abrufen von früheren Lernergebnissen, die zum Teil als eine Funktion von Gf generiert wurden (Cattell, 1963). Gc wurde dabei nicht nur als eine Funktion von Gf betrachtet, sondern auch als ein Produkt von Gelegenheit zum Erwerb, Persönlichkeitseigenschaften und Motivation (Cattell, 1963). Ein wesentlicher Unterschied zwischen den beiden Faktoren besteht hinsichtlich des vorhergesagten Entwicklungsverlaufs: Es wurde angenommen, dass es sich bei Gf um einen relativ stabilen und im großen Maße biologisch determinierten Faktor handelt, während Gc durch Erfahrung beeinflusst wird und über die Lebensspanne ansteigt (Horn & Cattell, 1966, 1967).

Die Gf-Gc Theorie stellt ein hierarchisches Modell dar. Die beiden Faktoren Gf und Gc wurden als Faktoren zweiter Ordnung betrachtet, denen eine Reihe von enger gefassten Primär-Fähigkeiten untergeordnet sind. Beispielsweise wird angenommen, dass der Faktor Gf Fähigkeiten wie *associative memory* oder *semantic relations* beeinflusst, während Gc den Fähigkeiten *verbal comprehension* oder *mechanical knowledge* übergeordnet ist (siehe Horn

& Cattell, 1966). Faktoranalytische Studien sowie Untersuchungen der angenommenen differentiellen Entwicklung von Gf und Gc lieferten deutliche Evidenz für die Existenz der beiden Faktoren (Horn & Cattell, 1967). In späteren Verfeinerungen des Modells (Hakstian & Cattell, 1978; Horn & Cattell, 1966), wurden eine Reihe weiterer Faktoren vorgeschlagen, die auf gleicher Ebene der Generalität wie Gf und Gc angesiedelt wurden. Zu diesen gehören beispielsweise General Visualization Capacity (Gv), General Speediness (Gs) sowie General Retrieval Capacity (Gr).

EI wurde häufig als eine Form von Gc betrachtet (Davies, Stankov & Roberts, 1998; MacCann, 2010; Zeidner, Roberts & Matthews, 2008). Diese Schlussfolgerung beruhte auf Studien, die den Zusammenhang zwischen EI und Maßen für Gf und Gc untersucht haben. Die beiden Operationalisierung des Four-Branch Modells, MEIS und MSCEIT, wiesen höhere Korrelationen mit Gc als mit Gf auf (siehe Roberts et al., 2008). EU wies über mehrere Studien hinweg von allen Branches die höchste Beziehung mit Gc auf (.40). Belege für die hieraus abgeleitete Annahme, dass sich EU durch ein Training erhöhen lässt, konnten bislang noch nicht gefunden werden (Nelis et al., 2009). Farrelly und Austin (2007) wiesen darauf hin, dass die bisherige Schlussfolgerung, EI sei eine Form von Gc, allein auf den bisherigen wenigen Operationalisierungen beruht. Es wäre denkbar, dass für EI auch eine fluide Komponente gefunden wird, wenn neue Instrumente entwickelt werden (Farrelly & Austin, 2007).

### **2.4.3 Facetten-Modelle der Intelligenz**

*Facetten-Modelle* der Intelligenz stellen eine Gruppe von Modellen dar, in denen intelligentes Verhalten durch verschiedene Merkmale klassifiziert werden kann, die als Facetten bezeichnet werden (Süß & Beauducel, 2005). Zu Facetten zählen beispielsweise die notwendigen Operationen, die bei der Bearbeitung der Testaufgabe notwendig sind (wie Schlussfolgern oder Merken) oder die Inhalte der Aufgabe (wie verbales oder numerisches

Material). Die Begriffe Facette und Facetten-Theorie wurden 1954 von Guttman vorgeschlagen und sind auf verschiedene inhaltliche Domänen der psychologischen Forschung angewendet worden (siehe Guttman & Greenbaum, 1998). Facetten bieten allerdings nicht nur die Möglichkeit der systematischen Beschreibung von Tests, sondern liefern auch prüfbare Hypothesen darüber, wie Testaufgaben miteinander zusammenhängen (Süß & Beauducel, 2005). Dabei wird angenommen, dass Tests umso höher miteinander korreliert sind, je ähnlicher sie sich hinsichtlich ihrer Ausprägungen auf den Facetten sind (Guttman, 1959).

Ein prominentes Facetten-Modell ist das Structure-of-Intellect Modell (SOI-Modell) von Guilford (siehe Guilford, 1959). Im SOI wurden die drei Facetten *Operationen*, *Inhalte* und *Produkte* vorgeschlagen, die jeweils mehrere Elemente umfassen (Guilford, 1959). Guilford ordnete sein Modell in Form eines Würfels an, dessen drei Dimensionen jeweils eine der Facetten darstellen. Aus den Kombinationen der drei Facetten (dargestellt als Zellen im Würfel) ergeben sich die vorgeschlagenen Fähigkeiten (siehe Guilford, 1959). Guilford schlug mehrere Revisionen seines ursprünglichen Modells vor, bei dem vor allem die Anzahl der Elemente der Facetten variierte (siehe Guilford 1966; 1988). Guilford (1988) nahm an, dass sich Intelligenz anhand von 6 Operationen (Evaluation, Convergent Production, Divergent Production, Memory Retention, Memory Recording, Cognition), 5 Inhalten (Visual, Auditory, Symbolic, Semantic und Behavioral) und 6 Produkten (Units, Classes, Relations, Systems, Transformations, Implications) beschreiben lässt. Insgesamt ergaben sich somit 180 Fähigkeiten als Kombinationen aus den drei Facetten.

Das im SOI-Modell vorgeschlagene Element Behavioral auf der Facette Inhalte ist für die EI besonders interessant. Der Inhalt Behavioral bezieht sich auf das Verhalten anderer Personen. O`Sullivan und Guilford (1975) konnten Belege für 6 Faktoren identifizieren, die sie theoretisch aus dem SOI-Modell als Kombinationen des Inhaltes Behavioral und der

Operation Cognition ableiteten. O`Sullivan und Guilford (1975) schrieben diese Faktoren der sozialen Intelligenz zu und definierten sie als Fähigkeit „to understand other people’s thoughts, feelings, and intentions“ (O`Sullivan & Guilford, 1975, S. 256). Da sich diese Fähigkeit auch auf Emotionen anderer Personen bezieht, ist die Facette Behavioral für die EI von Relevanz. In ihren Analysen zogen O`Sullivan und Guilford (1975) eine Reihe von Tests heran, die die Bedeutung für EI und für EU illustrieren. Beispielsweise wurde als ein Indikator für die Facetten-Kombination *Cognition, Behavioral und Systems* der Test *Facial Situations* herangezogen, bei dem Testpersonen aus vorgegebenen Situationen diejenige wählen soll, die einen vorgegebenen Emotionsausdruck hervorgerufen haben könnte (O`Sullivan & Guilford, 1975).

Im Rahmen des SOI-Modells könnte EI einen Teil der Fähigkeiten darstellen, die sich aus den Kombinationen des Inhaltes Behavioral mit den anderen Facetten ergeben. Es wäre darüber hinaus auch denkbar, dass man EI in Facetten-Modelle integriert, indem man sie um einen Inhalt der Emotion erweitert.

#### **2.4.4 Soziale Intelligenz**

Soziale Intelligenz ist schon früh in der Geschichte der Intelligenz-Theorien als eine wichtige menschliche Fähigkeit vorgeschlagen worden. Thorndike (1920) argumentierte, dass Personen sich nicht nur hinsichtlich einer einzelnen Intelligenz beschreiben lassen, sondern ein Muster von Ausprägungen auf verschiedenen Intelligenzformen aufweisen. Thorndike schlug drei Arten der Intelligenz vor, die aus seiner Sicht in vielen Fällen ausreichend sind, um Personen beschreiben zu können: Mechanische Intelligenz, Abstrakte Intelligenz und Soziale Intelligenz. Soziale Intelligenz wurde von Thorndike definiert als “the ability to understand and manage men and women, boys and girls – to act wisely in human relations” (Thorndike, 1920, S. 228). Einen der wichtigsten Ansätze für die Konzeptualisierung und Messung Sozialer Intelligenz lieferte Guilford im Rahmen des SOI-Modells (siehe auch

O'Sullivan & Guilford, 1975; Weis & Süß, 2005). Trotz der relativ langen Geschichte der Sozialen Intelligenz, sind bis heute einige entscheidende Probleme der Theorie und der Messung ungelöst (Weis & Süß, 2005). Ein Problem besteht darin, dass sich Soziale Intelligenz empirisch kaum von akademischer und vor allem verbaler Intelligenz trennen lässt (Kang, Day & Meara, 2005; Riggio, Messamer & Throckmorton, 1991). Andere konnten hingegen erste Evidenz für einen Faktor der sozialen Intelligenz finden (z. B. Ford & Tisak, 1983; Weis & Süß, 2007). Insgesamt wird die Suche nach Sozialer Intelligenz häufig pessimistisch betrachtet (siehe Landy, 2006).

Glücklicherweise gibt es neuere Entwicklungen im Bereich der Sozialen Intelligenz, die recht vielversprechend erscheinen. Weis und Süß (2005) schlugen auf Grundlage bisheriger Erkenntnisse aus der Forschung zur Sozialen Intelligenz ein Modell vor, in dem fünf Facetten der Sozialen Intelligenz angenommen werden: Social Understanding, Social Memory, Social Perception, Social Creativity und Social Knowledge. Für einige der vorgeschlagenen Fähigkeiten konnte erste Evidenz gefunden werden (siehe Conzelmann, Weis & Süß, 2013; Weis & Süß, 2007).

Bereits bei Einführung des Konstruktes EI machten Salovey und Mayer (1990) auf eine Überschneidung mit Sozialer Intelligenz aufmerksam. EI wurde als der Teil der sozialen Intelligenz betrachtet, der sich auf emotionsbezogene Probleme bezieht (Mayer et al., 1999). EI sei andererseits aber auch breiter als Soziale Intelligenz, da sich EI nicht nur auf die Emotionen Anderer, sondern auch die eigenen Emotionen bezieht (Mayer et al., 1999). Besonders die Facette Social Understanding könnte für die Domäne EI sehr interessant sein (Weis & Süß, 2007). Weis und Süß (2007) definierten Social Understanding als Fähigkeit, „to understand or interpret given social stimuli in terms of the implications for the person or situation“ (S. 5). Operationalisierungen dieser Fähigkeiten umfassen u. a. das Wahrnehmen von Emotionen anderer Personen oder das Erschließen von Emotionen anhand von Hinweisen

über die Situation (Conzelmann et al., 2013). Hier wird eine Überschneidung mit EI und insbesondere mit EU deutlich.

#### **2.4.5 Gardners Theorie der Multiplen Intelligenzen**

In Gardners Theorie der Multiplen Intelligenzen wird angenommen, dass es mindestens acht relativ unabhängige Formen der Intelligenz gibt (Davidson & Downing, 2000; Gardner & Hatch, 1989). Diese Formen der Intelligenz sind die linguistische, logisch-mathematische, räumliche, musikalisch, körperlich-kinästhetisch, intrapersonale, interpersonale und naturalistische Intelligenz (Davidson & Downing, 2000). Als Begründungen für die Theorie der Multiplen Intelligenz wies Gardner (1991) auf verschiedene Beobachtungen hin, wie beispielsweise darauf, dass bei Hirnschädigungen ein Teil von Fähigkeiten betroffen sein kann, während andere Fähigkeiten unbeeinflusst sind, oder dass Individuen hohe Begabungen in bestimmten Domäne aufweisen können, während andere Bereiche normal oder unterdurchschnittlich ausgeprägt sind.

Im Rahmen dieser Theorien, haben zwei Formen der Intelligenz, die später auch als eine Form betrachtet wurden, theoretische Überschneidungen mit EI (Gardner & Hatch, 1989; Waterhouse, 2006): *Intrapersonale* und *Interpersonale Intelligenz*. Interpersonale Intelligenz wurde als Fähigkeit definiert, auf die Stimmungen, das Temperament, Motivationen und Bedürfnisse anderer Menschen eingehen zu können (Gardner & Hatch, 1989). Intrapersonale Intelligenz beinhaltet die Fähigkeit, Zugang zu eigenen Gefühlen zu haben, sie unterscheiden und nutzen zu können sowie als Wissen über eigene Stärken, Schwächen und Bedürfnisse (Gardner & Hatch, 1989). Für die Theorie der Multiplen Intelligenzen gibt es wenige Belege und einige ihrer Grundannahmen wurden kritisiert (siehe Klein, 1997; Waterhouse, 2006).

## 2.5 EI und Persönlichkeitseigenschaften

Persönlichkeitseigenschaften können als stabile, dispositionelle Charakteristika einer Person definiert werden, die das Verhalten über verschiedene Situationen hinweg beeinflussen (Schulze et al., 2005, S. 16). Im Sinne der Taxonomie von Cronbach, sind Persönlichkeitseigenschaften wichtige Determinanten des typischen Verhaltens einer Person. Theorien über Persönlichkeitseigenschaften haben eine lange Geschichte, in der verschiedene Modelle vorgeschlagen wurden (John & Srivastava, 1999). Diese Modelle unterscheiden sich hinsichtlich der Anzahl und Art der angenommenen Dimensionen zur Beschreibung von sogenannten “emotional, interpersonal, experiential, attitudinal, and motivational styles” (McCrae & John, 1992, S. 175). Von den bisher vorgeschlagenen Modellen stellt das *Fünf-Faktoren Modell der Persönlichkeit* das mittlerweile am meisten akzeptierte Modell dar (John, Nauman & Soto, 2008), das allerdings nicht unumstritten ist (siehe Eysenck, 1992). Im Fünf-Faktoren Modell werden Persönlichkeitseigenschaften als hierarchische Taxonomie aufgefasst, in der fünf grundlegende Dimensionen der Persönlichkeit angenommen werden: *Extraversion, Neurotizismus, Gewissenhaftigkeit, Verträglichkeit* sowie *Offenheit für Erfahrungen* (z. B. McCrae & John, 1992). Diese fünf Faktoren sind auch unter dem Begriff der *Big-Five* bekannt. Nach Costa und McCrae (1992) soll die Bezeichnung Big-Five nicht implizieren, dass im Fünf-Faktoren Modell die Existenz von nur ausschließlich fünf wichtigen Persönlichkeitseigenschaften behauptet wird. Stattdessen kann das Fünf-Faktoren Modell als ein hierarchisches Modell verstanden werden, in denen die Big-Five sehr globale Beschreibungen der Persönlichkeit darstellen, denen mehrere, spezifische Persönlichkeitseigenschaften untergeordnet sind (Costa & McCrae, 1992; McCrae & John, 1992). Costa und McCrae (1992) schlugen 30 untergeordnete Persönlichkeitseigenschaften vor, die sie in ihrem Modell als Facetten bezeichneten.

Das Fünf-Faktoren Modell wurde aus Analysen von persönlichkeitsbeschreibenden Begriffen der Sprache abgeleitet, was auch als *lexikaler Ansatz* bezeichnet wird (siehe z. B. John et al., 2008). Die dem lexikalen Ansatz zugrundeliegende sogenannte *lexikale Hypothese* besagt, dass in der Sprache sozial relevante Persönlichkeitseigenschaften enthalten sind, da Sprache u. a. benutzt wird, um sich und andere zu beschreiben (John & Srivastava, 1999). Die hohe Akzeptanz des Modells beruht auf einer großen Anzahl von Studien, die Evidenz für die im Modell getroffenen Annahmen erbracht haben. Es wurde gezeigt, dass die fünf Faktoren robust in unterschiedlichen Sets von Eigenschaftsbegriffen und Instrumenten zur Persönlichkeitsbeschreibung gefunden werden können (Digman & Inouye, 1986; Goldberg, 1990). Des Weiteren sind die Big-Five universell in unterschiedlichen Kulturen oder Altersbereichen zu finden und scheinen eine biologische Basis zu haben (Costa & McCrae, 1992).

Intelligenz hat sich als relativ unabhängig von Persönlichkeitsfaktoren erwiesen (siehe Ackerman & Heggestad, 1997). In einer Meta-Analyse von Ackerman und Heggestad (1997) wurden u. a. die Beziehungen der Big-Five zu Allgemeiner Intelligenz, Gf und Gc untersucht. Extraversion, Verträglichkeit, Gewissenhaftigkeit und Neurotizismus wiesen zu Intelligenz entweder keine Beziehungen oder Beziehungen nahe Null auf. Für Offenheit für Erfahrungen wurden allerdings moderate positive Beziehungen mit Allgemeiner Intelligenz sowie Gc gefunden. Als Form von Intelligenz sollte EI ebenfalls ein Muster an Beziehungen zu den Big-Five aufweisen, das mit klassischer Intelligenz vergleichbar ist. Für neu entwickelte EI-Tests erscheint eine Überprüfung dieses Kriterium wichtig zu sein, angesichts der vielen Modelle und Instrumente unter dem Label EI, die diesem Kriterium nicht standhalten. Wie theoretisch zu erwarten weisen Trait-EI und Mixed Modells teilweise hohe Korrelationen mit den Big-Five auf (siehe Roberts et al., 2008). Korrelationen sind daneben schon alleine deshalb zu erwarten, da bei der Konstruktion von Instrumenten für Mixed Models und Trait-EI oft Material aus etablierten Persönlichkeitstests übernommen wurde (Wilhelm, 2005). Für



Operationalisierungen des Four-Branch Modells hingegen konnte gezeigt werden, dass die angenommenen Fähigkeiten weitestgehend unabhängig von den Big-Five sind (Roberts et al., 2008). Roberts et al. (2008) fanden geringe positive Beziehungen für Facilitation, Understanding und Management mit Offenheit für Erfahrungen. Darüber hinaus wurden auch geringe positive Beziehungen der vier Branches zu Verträglichkeit gefunden, wobei hier Management die höchste Beziehung aufwies. Insgesamt weisen fähigkeitsbasierte Ansätze (siehe auch MacCann & Roberts, 2008) Muster an Beziehungen zu den Big-Five auf, die weitestgehend den Erwartungen für eine Form von Intelligenz entsprechen.

## **2.6 EU außerhalb der Theorie der EI**

EU ist mit zwei Konzepten eng verknüpft, die außerhalb der Forschung zur EI behandelt werden: *Affective Forecasting-Fähigkeit* und *Empathie*. Da sich Affective Forecasting-Fähigkeit auf das Verstehen der eigenen Emotionen und Empathie auf das Verstehen der Emotionen anderer Personen bezieht, ist eine Betrachtung dieser beiden Konzepte im Zusammenhang mit EU wichtig.

### **2.6.1 Affective Forecasting-Fähigkeit**

Der Begriff Affective Forecasting bezeichnet die Vorhersage einer Person über ihre eigenen zukünftigen Emotionen (Wilson & Gilbert, 2003). Um die Genauigkeit von Affective Forecasts zu erfassen, wird typischerweise die Übereinstimmung von vorhergesagten Emotionen vor einem Ereignis mit den selbstberichteten Emotionen nach dem Ereignis ermittelt. Ein Großteil der Forschung zu Affective Forecasting thematisiert die systematischen Fehler, die Personen bei der Vorhersage ihrer eigenen Emotionen unterlaufen. In diesem Zusammenhang wurde die Genauigkeit verschiedener Aspekte der Vorhersage betrachtet, wie die Valenz, Intensität und Dauer der vorhergesagten Emotionen (Wilson & Gilbert, 2003). Beispielsweise wurde festgestellt, dass Personen die Dauer (*durability bias*) sowie die Intensität (*impact bias*) ihrer Emotionen für ein zukünftiges Ereignis systematisch

überschätzen (siehe Wilson & Gilbert, 2003). Eine Ursache für Verzerrungen dieser Art könnte darin liegen, dass Personen bei ihren Vorhersagen zu wenig den Einfluss von weiteren Faktoren berücksichtigen, die in der Zukunft, neben dem Ereignis selbst, auf den emotionalen Zustand wirken können (Wilson, Meyers & Gilbert, 2001). Unter anderem könnten Personen dem Effekt der Emotionsregulation auf die Intensität und Dauer ihrer Emotionen nicht genügend Rechnung tragen (Van Dijk, Van Dillen, Seip & Rotteveel, 2012).

Hinsichtlich der Genauigkeit der Vorhersage eigener Emotionen wurde zudem die Rolle von vorherigen Erfahrungen mit einem vergleichbaren Ereignis diskutiert. In einer Reihe von Studien konnte gezeigt werden, dass Personen bei der Vorhersage zukünftiger Emotionen auf Erfahrungen mit ähnlichen Ereignissen zurückgreifen und ihre Vorhersagen daran anpassen (Hoerger, Chapman, Epstein & Duberstein, 2012; Meyvis, Ratner & Levav, 2010; Wilson et al., 2001). Allerdings ließen sich Fehler bei den Vorhersagen, anders als erwartet, nicht immer durch vorherige Erfahrungen reduzieren (Wilson et al., 2001). Wilson et al. (2001) erklärten die Konsistenz des Auftretens systematischer Verzerrungen damit, dass Personen in vielen Situationen nicht motiviert und/oder in der Lage sind, ihre Vorhersagen an Erfahrungen mit vergleichbaren Ereignissen zu adaptieren. Fehlerhafte Vorhersagen könnten trotz Erfahrungen auftreten, wenn Personen a) ihre Erfahrungen mit vergangenen Ereignissen bei der Vorhersage nicht genügend miteinbeziehen, b) nicht korrekt einschätzen, welche Erfahrungen mit dem zukünftigen Ereignis vergleichbar sind und damit bei der Vorhersage anwendbar sind oder c) nicht korrekt abrufen, wie sie sich in der Vergangenheit gefühlt haben (Wilson et al., 2001).

Neben der Betrachtung von systematischen Verzerrungen, die typischerweise bei der Vorhersage eigener Emotionen auftreten, wurden auch individuelle Unterschiede bezüglich der Fähigkeit, akkurate Vorhersagen zu treffen, untersucht. Die sogenannte Affective Forecasting-Fähigkeit wurde von einigen Autorinnen und Autoren mit EI und insbesondere

mit EU in Verbindung gebracht (Hoerger et al., 2012; Dunn, Brackett, Ashton-James, Scheiderman & Salovey, 2007; Orchard et al., 2009). Hoerger et al. (2012) wiesen darauf hin, dass es der Forschung zu individuellen Unterschieden in der Genauigkeit von Affective Forecasts bisher an einem theoretischen Rahmen fehlt und dass EI diesen liefern könnte. EI könnte einen Einfluss auf die Akkuratheit von Affective Forecasts haben, indem sie beeinflusst, wie gut Personen sich episodisches Wissen über vergangene emotionale Erlebnisse aneignen und diese bei der Vorhersage abrufen können (Hoerger et al., 2012). Für die theoretisch erwarteten Zusammenhänge von Affective Forecasting-Fähigkeit mit EI und EU liegen erste Belege vor. Zum einen weisen Maße für EI (Dunn et al., 2007) und EU (Hoerger et al., 2012) die angenommenen positiven Zusammenhänge mit Affective Forecasting-Fähigkeit auf. Der positive Zusammenhang zwischen EU und Affective Forecasting-Fähigkeit blieb in einer Studie von Hoerger et al. (2012) selbst dann bestehen, wenn kognitive Fähigkeiten statistisch kontrolliert wurden. Zum anderen lieferten Hoerger et al. (2012) direkte Hinweise über die Rolle, die EU beim Affective Forecasting spielen könnte. Hierzu wurde die Vorhersagegenauigkeit von Emotionen für zukünftige Ereignisse in zwei Phasen miteinander verglichen: einmal bevor und einmal nachdem die Untersuchungsteilnehmerinnen und -teilnehmer Erfahrungen mit einem vergleichbaren Ereignis gemacht hatten. Mit EU konnten Verbesserungen der Akkuratheit von Affective Forecasts als Resultat von Erfahrungen mit einem vorangegangenen, vergleichbaren Ereignis vorhergesagt werden (Hoerger et al., 2012). Darüber hinaus konnte der Zusammenhang zwischen EU und Verbesserungen der Akkuratheit von Affective Forecasts teilweise über eine bessere Erinnerungsleistung an die eigenen emotionalen Reaktionen, gemessen kurze Zeit nach dem Ereignis, erklärt werden (Hoerger et al., 2012). Damit könnte EU eine Rolle dabei spielen, wie gut Personen in der Lage sind, Wissen aus vergangenen Erfahrungen zu erwerben und dieses bei der Vorhersage zukünftiger Ereignisse anzuwenden.

Trotz dieser interessanten Befunde wurde Affective Forecasting-Fähigkeit bisher äußerst selten im Rahmen der EI diskutiert und es gibt nur wenige Studien, die den angenommenen Zusammenhang zwischen EI und Affective Forecasting-Fähigkeit untersucht haben (Dunn et al., 2007; Hoerger et al., 2012). Orchard et al. (2009) machten darauf aufmerksam, dass Paradigmen, die zur Erfassung von Affective Forecasting-Fähigkeit genutzt werden, auch zur Weiterentwicklung der Messung von EU geeignet sind oder zur Validierung von Tests der EI herangezogen werden könnten.

### **2.6.2 Empathie**

Im Alltagsverständnis wird Empathie als eine wichtige Eigenschaft von Individuen betrachtet, die bspw. von Bewerberinnen und Bewerbern für verschiedene Berufe gefordert wird oder als ein Ziel von Fördermaßnahmen im schulischen Kontext genannt wird. Empathie wird hier als Bereitschaft und Fähigkeit aufgefasst, sich in andere Personen einzufühlen („Empathie“, n.d.). Auch in der wissenschaftlichen Domäne wurde Empathie als eine wichtige Fähigkeit vorgeschlagen, von der u. a. angenommen wird, dass sie zu einer Minderung aggressiven Verhaltens beiträgt sowie prosoziales Verhalten fördern kann (siehe Davis, Mitchell, Hall, Lothert, Snapp & Meyer, 1999; Vachon, Lynam & Johnson, 2014). Ein Mangel an empathischen Fähigkeiten wurde darüber hinaus mit verschiedenen psychischen Störungen in Verbindung gebracht, wie bspw. Autismus oder Schizophrenie (Baron-Cohen & Wheelwright, 2004; Michaels, Horan, Ginger, Martinovich, Pinkham & Smith, 2014). Geringe empathische Fähigkeiten wurden in diesem Zusammenhang als eine Ursache für Einschränkungen der sozialen Funktionsfähigkeit der Betroffenen betrachtet (Michaels et al., 2014). Insbesondere im klinischen Kontext wurde für das Merkmal gering ausgeprägter emotionaler Fähigkeiten auch das Konzept der *Alexithymie* vorgeschlagen, das u. a. durch einen Mangel an Empathie-Fähigkeit charakterisiert ist (Grynberg, Luminet, Corneille, Grèzes & Berthoz, 2010).

In der wissenschaftlichen Literatur gibt es keine universell akzeptierte Definition der Empathie (Bernhardt & Singer, 2012). Empathie wird zumeist als eine bestimmte Form der Perspektivübernahmefähigkeit betrachtet, die es ermöglicht, sich in die Emotionen einer anderen Person hineinzusetzen (siehe z. B. Niedenthal et al., 2006; Ruby & Decety, 2004). In einigen Definitionen wird das Teilen des affektiven Zustandes mit einer Zielperson als ein wesentlicher Bestandteil der Empathie betrachtet (z. B. Bernhardt & Singer, 2012; Fan, Duncan, de Greck & Northoff, 2011). Beispielsweise definierten Bernhardt und Singer (2012) Empathie als „... ability to share the affective states of our closest ones and complete strangers...“ (S. 2), die es ermöglicht „... to predict and understand their feelings, motivations, and actions.“ (S. 2). Im Rahmen der Empathie-Forschung wurden eine Reihe verwandter Konstrukte vorgeschlagen, von der Empathie abgegrenzt werden kann. Ein Beispiel hierfür ist das Konzept des *Emotional Contagion*. Emotional Contagion umfasst ebenfalls das Teilen eines affektiven Zustandes mit einer Zielperson, unterscheidet sich allerdings hinsichtlich eines entscheidenden Aspektes von Empathie: Im Prozess der Empathie löst die Vorstellung des affektiven Zustandes der Zielperson den eigenen affektiven Zustand hervor (Bernhardt & Singer, 2012). Beim Emotional Contagion hingegen übernimmt das Individuum automatisch den affektiven Zustand einer Zielperson, ohne dass die Ursache für den eigenen affektiven Zustand der Zielperson zugeschrieben wird (Bernhardt & Singer, 2012; Fan et al., 2011). Das Merkmal des Teilens des affektiven Zustands unterscheidet Empathie zudem von verwandten Konzepten, die oft als Auslöser prosozialen Verhaltens gesehen werden, wie *Sympathy* oder *Empathic Concern* (siehe Bernhardt & Singer, 2012). Unter *Sympathy* werden Emotionen gegenüber einer Zielperson verstanden, die zwar durch deren affektiven Zustand ausgelöst werden, allerdings nicht der Perspektivübernahme dienen, wie z. B. die Besorgnis um die Zielperson. Unter *Empathic Concern* wird der motivationale Zustand verstanden, einer Zielperson helfen zu wollen, der auf die Perspektivübernahme folgen kann (Bernhardt & Singer, 2012).

Im Rahmen von Ansätzen, in denen das Teilen des affektiven Zustandes als ein wesentlicher Bestandteil der Empathie betrachtet wird, wird das kognitive Verstehen von Emotionen häufig begrifflich vom Konzept der Empathie abgegrenzt. Für das Verstehen von Emotionen wurden Begriffe eingeführt, wie z. B. *Affective Theory of Mind* oder *Empathic Accuracy* (z. B. Ickes, 2001; Van Boven & Loewenstein, 2001).

In anderen Ansätzen wurde angenommen, dass Empathie sowohl das Nachempfinden als auch das kognitive Verstehen von Emotionen beinhalten. Davis (1983) schlug vor, dass Empathie zwei Komponenten umfasst: *kognitive Empathie* (d. h. das Verstehen von Emotionen) und *affektive Empathie* (d. h. das Nachempfinden von Emotionen). Über diese Annahme besteht in der derzeitigen Literatur zur Empathie ein gewisser Konsens (siehe z. B. Zaki, Bolger & Ochsner, 2008). Einige Autorinnen und Autoren schlugen vor, dass affektive Empathie eine Grundlage für das Verstehen von Emotionen liefern kann, indem die Emotionen der Zielperson zum besseren Verständnis simuliert werden (Batson, 2009; Bernhardt & Singer, 2012; Zaki et al., 2008). In Bezug auf das Simulieren von Emotionen, wurden zudem zwei mögliche Prozesse voneinander unterschieden: neben dem Versuch, die Emotionen der Zielperson nachzuempfinden, könnten Personen auch versuchen, ihre eigenen Emotionen zu simulieren, die wahrscheinlich in der gegebenen Situation auftreten werden (Batson, 2009). Letzteres entspricht in gewisser Weise einem Affective Forecast, der im Sinne der Perspektivübernahme zur Vorhersage der Emotionen einer anderen Person herangezogen wird. Das Simulieren der eigenen Emotionen könnte vor allem dann eine Basis für das Verstehen von Emotionen liefern, wenn die Zielperson und/oder die Situation wenig bekannt sind (Batson, 2009). Andererseits wurde das Vorstellen der eigenen Emotionen zur Perspektivübernahme einer anderen Person auch als eine mögliche Fehlerquelle betrachtet (Batson, 2009; Ruby & Decety, 2004; Van Boven & Loewenstein, 2005). Das Simulieren der eigenen Emotionen könnte insbesondere dann zu falschen Vorhersagen über die Emotionen

der Zielperson führen, wenn sich diese hinsichtlich relevanter Persönlichkeitseigenschaften von der eigenen Person unterscheidet (Batson, 2009).

Ein großer Teil der Forschung zur Empathie beschäftigt sich mit der Identifizierung einer neurologischen Basis für den Prozess der Perspektivübernahme und insbesondere für das Nachempfinden von Emotionen. In einigen Studien wurden Belege dafür erbracht, dass bei der Beobachtung und beim Vorhersagen von Emotionen anderer Personen bestimmte Regionen des Gehirns aktiviert sind, die auch am direkten Erleben von Emotionen beteiligt sind (siehe Bernhardt & Singer, 2012). In einer Meta-Analyse waren Aktivierungen dieser Regionen und das Nachempfinden verschiedenster Emotionen konsistent miteinander assoziiert (Fan et al., 2011). Fan et al. (2011) fanden zudem Hinweise darauf, dass beim Verstehen und Nachempfinden von Emotionen ähnliche, aber dennoch unterscheidbare Hirnregionen aktiviert sind, was aus Sicht der Autorinnen und Autoren einen Hinweis darauf liefert, dass es sich bei kognitiver und affektiver Empathie tatsächlich um trennbare Prozesse handeln könnte. Des Weiteren wurde vorgeschlagen, dass eine bestimmte Klasse von Neuronen, die sogenannten Spiegelneurone, am Prozess der Empathie beteiligt sind (z. B. Iacoboni, 2009; Schulte-Rüther, Markowitsch, Fink & Piefke, 2007). Spiegelneurone sind dadurch charakterisiert, dass sie sowohl bei der Ausführung einer Handlung als auch bei der Beobachtung der gleichen Handlung aktiviert sind (Iacoboni, 2009). Analog dazu könnten Spiegelneurone auch beim Nachempfinden von Emotionen involviert sein (z. B. Iacoboni, 2009). Die angenommene Rolle von Spiegelneuronen im Prozess der Empathie ist allerdings nicht unumstritten und die Forschung hierzu befindet sich noch in einer eher frühen Phase (siehe Decety, 2010; Fan et al., 2011).

Hinsichtlich der Konzeptualisierung und Messung von Empathie kann eine Zweiteilung beobachtet werden, die derjenigen im Bereich der EI sehr ähnlich ist. Zum einen wurde Empathie als ein Fähigkeitskonstrukt definiert und es gab Bemühungen,

leistungsbasierte Tests sowie geeignete Scoring-Mechanismen zu entwickeln, die zum Teil weit in die Empathie-Forschung zurückreichen (Cronbach, 1955; Dymond, 1949; Dziobek, Rogers, Fleck, Bahnemann, Heekeren, Wolf & Convit, 2008; Ickes, 2001). In leistungsbasierten Tests der Empathie besteht die Aufgabe der Testperson typischerweise darin, die Emotionen einer Zielperson in einer bestimmten Situation zu erschließen und vorherzusagen (siehe z. B. Ickes, 2001). Als Kriterium zur Bewertung, ob die Emotionen der Zielpersonen korrekt von der Testperson verstanden wurden, werden Selbstauskünfte der Zielperson oder Einschätzungen von Expertinnen und Experten herangezogen (Cronbach, 1955; Dziobek et al., 2008).

Zum anderen wurde Empathie häufig als eine Mischung aus Persönlichkeitseigenschaften und Fähigkeiten konzeptualisiert. In diesem Zusammenhang wurden eine Reihe von Selbstberichten entwickelt, wie dem Interpersonal Reactivity Index (IRI; Davis, 1983), dem Empathy Quotient (EQ; Baron-Cohen & Wheelwright, 2004) sowie der Toronto Empathy Questionnaire (TEQ; Spreng, McKinnon, Mar & Levine, 2009). Im IRI von Davis (1983) wurden bspw. vier verschiedene Konstrukte vorgeschlagen, die jeweils Teilaspekte der kognitiven und affektiven Empathie widerspiegeln sollen: Perspective-Taking, Fantasy, Empathic Concern und Personal Distress (Davis, 1983). Im IRI wird Empathie als Verhaltenstendenz operationalisiert und mit Items wie „I sometimes try to understand my friends better by imagining how things look from their perspective“ (Perspective-Taking) oder „I often have tender, concerned feelings for people less fortunate than me“ (Empathic Concern) erfasst (Davis, 1983).

Die Unterscheidung von Empathie als Fähigkeit oder Persönlichkeitseigenschaft und die damit verknüpfte Messung in Form von Tests maximalen oder typischen Verhaltens ist zudem mit dem Konzept der kognitiven und affektiven Empathie assoziiert. Häufig wird kognitive Empathie eher als eine Fähigkeit und affektive Empathie eher als eine Disposition



aufgefasst (z. B. Zaki et al., 2008). Während kognitive Empathie, neben Selbsteinschätzungen, auch über Leistungsmaße erfasst wird (z. B. Ickes, 2001), wird affektive Empathie zumeist über Selbstberichte erfasst (Losoya & Eisenberg, 2001). Neben Selbstberichten zur Messung von affektiver Empathie wurden auch physiologische Maße vorgeschlagen, wie die Herzrate oder die Hautleitfähigkeit, die weniger im Verdacht stehen, von der Testperson verfälschbar zu sein (Losoya & Eisenberg, 2001). Diese Art der Messung bringt allerdings auch Probleme mit sich. Die gemessenen physiologischen Reaktionen sind nicht für eine bestimmte Emotion spezifisch und können in der Testsituation zudem leicht durch andere Faktoren als Emotionen beeinflusst worden sein (siehe Losoya & Eisenberg, 2001).

Ein Beispiel für einen Test, der sowohl kognitive als auch affektive Empathie standardisiert erfassen soll, stellt der Multifaceted Empathy Test (MET; Dziobek et al., 2008) dar, der für Erwachsene mit Autismus entwickelt wurde. Im MET werden den Testpersonen Fotos präsentiert, auf denen eine Zielperson in einem bestimmten Kontext dargestellt ist. Zur Erfassung kognitiver Empathie sollen die Testpersonen die Emotionen der Zielperson erschließen. Die Korrektheit der Testantworten wird im MET durch Expertenurteile festgelegt. Zur Messung affektiver Empathie werden die Testpersonen anschließend gebeten, ihre eigenen emotionalen Reaktionen bei der Betrachtung der Fotos zu berichten. Auch in diesem Test ist die Erfassung von affektiver Empathie auf Selbstberichte der Testpersonen gestützt.

Trotz ihrer Ähnlichkeit wurden EU und Empathie-Fähigkeit bisher eher isoliert voneinander betrachtet. Es gibt nur wenige Studien, in denen empathische Fähigkeiten mit EU oder EI in Verbindung gebracht wurden. In diesen konnten aber erste Hinweise dafür erbracht werden, dass die erwarteten positiven Zusammenhänge zwischen EI bzw. EU und Empathie-Fähigkeiten tatsächlich beobachtet werden können (Ferguson & Austin, 2010; Qualter,

Barlow & Stylianou, 2011). Als Fähigkeit lässt sich EU theoretisch eher dem zuordnen, was unter kognitiver Empathie verstanden wird. Allerdings könnte auch der affektiven Komponente der Empathie eine Fähigkeit unterliegen. Wenn angenommen wird, dass das Simulieren von Emotionen eine Grundlage für das Verstehen liefert könnte (siehe z. B. Bernhardt & Singer, 2012), stellt sich die Frage, auf welcher Grundlage das Vorstellen der Emotionen geschehen kann. Einen Hinweis hierauf kann die Affective Forecasting-Forschung liefern, in der angenommen wird, dass man zukünftige Ereignisse auf eigenen Erfahrungen basieren kann. Anstatt die Vorhersage auf Affective Forecasts zu basieren, könnten Personen zudem auf erworbenes Wissen über die Eigenschaften der Zielperson zurückgreifen, was in vielen Fällen akkuratere Vorhersagen erlaubt (siehe Batson, 2009). Insgesamt kann festgehalten werden, dass aufgrund der hohen konzeptuellen Überschneidungen zwischen Affective Forecasting-Fähigkeit, Empathie-Fähigkeit und EU, diese in der Zukunft häufiger gemeinsam betrachtet werden sollten. Es bleibt zudem zu klären, inwiefern es sich hierbei um drei unterscheidbare Konstrukte handelt oder ob nicht vielmehr eine gemeinsame Fähigkeit dahinter steht. Erkenntnisse über den Prozess der Empathie sowie der Affective Forecasting-Fähigkeit können sowohl für die Weiterentwicklung der Theorie als auch der Messung von EU hilfreich sein.

## **2.7 Theorien über den Emotionsauslösungsprozess**

In Theorien des Emotionsauslöseprozesses werden Regelmäßigkeiten beschrieben, die der Auslösung von Emotionen unterliegen. Ein wichtiger Teilaspekt dieser Theorien sind die verschiedenen Annahmen darüber, was Ursachen von Emotionen sind (siehe Moors, 2009). Für die Erforschung von EU sind Theorien des Emotionsauslösungsprozesses interessant, da sie Hinweise darüber liefern können, auf welcher Grundlage Personen Emotionen verstehen und vorhersagen können. Zudem wurden Theorien des Emotionsauslösungsprozesses als eine Basis für die Konstruktion und das Scoring von Tests zur Erfassung von EU diskutiert (siehe

MacCann & Roberts, 2008). Über den Emotionsauslösungsprozess gibt es eine Reihe von Theorien, die sich vor allem hinsichtlich der darin getroffenen Annahmen über die Abfolge des Prozesses sowie über die Rolle von Kognitionen bei der Emotionsauslösung unterscheiden (Niedenthal et al., 2006; Roseman & Smith, 2001).

### **2.7.1 Klassische Emotionstheorien**

In einigen klassischen Emotionstheorien wird davon ausgegangen, dass Emotionen durch bestimmte Reize ausgelöst werden, ohne dass Kognitionen dabei eine vermittelnde Rolle spielen (Roseman & Smith, 2001). Hierzu zählen evolutionäre Emotionstheorien, in denen angenommen wird, dass Emotionen eine biologische Basis haben (Niedenthal et al., 2006). Emotionen werden in diesem Zusammenhang als adaptive Reaktionen auf bestimmte Stimuli betrachtet, die sich im Laufe der Evolution entwickelt haben.

Einen weiteren Ansatz, in dem die Auslösung von Emotionen ohne kognitive Prozesse erklärt wird, stellen Reiz-Reaktions-Theorien dar. Watson und Rayner (1920) versuchten die Komplexität emotionaler Reaktionen auf verschiedene Reize sowie Unterschiede zwischen Individuen im Rahmen behavioristischer Lerntheorien zu erklären. Die Autoren zeigten, dass emotionale Reaktionen auf Reize ein Resultat von Konditionierung sein können (Watson & Rayner, 1920).

In der klassischen Emotionstheorie nach James (1884) spielen Kognitionen ebenfalls keine Rolle im Emotionsauslösungsprozess. Nach James stellen körperliche Reaktionen einen zentralen Aspekt der Auslösung von Emotion dar. In seiner Theorie schlug James einen Ablauf der Emotionsauslösung vor, der als revolutionär bezeichnet werden kann, da er den allgemeinen Vorstellungen und bisherigen Theorien über Emotionen widersprach (Moors, 2009). Typischerweise wurde angenommen, dass ein Stimulus zu einer Emotion (z. B. Angst) führt, die wiederum eine körperliche Reaktion auslöst (z. B. Fluchtreaktion). Nach der

Emotionstheorie von James löst ein Stimulus zunächst eine körperliche Reaktion aus, der erst anschließend die Wahrnehmung einer Emotion folgt.

James Theorie wurde sowohl hinsichtlich einiger theoretischer als auch empirischer Gesichtspunkte kritisiert (siehe Moors, 2009). Beispielsweise konnte die Theorie nicht erklären, wie die gleiche körperliche Reaktion (z. B. der Anstieg der Herzfrequenz) unterschiedliche Emotionen (z. B. Angst oder Ärger) hervorrufen kann (Cannon, 1927). Darüber hinaus gelang es in experimentellen Untersuchungen nicht, durch Auslösung von körperlichen Reaktionen Emotionen zu erzeugen, wie es im Rahmen der Theorie zu erwarten ist (Moors, 2009; Schachter & Singer, 1962).

Im Zusammenhang mit der Kritik an der klassischen Emotionstheorie nach James, wurden Theorien vorgeschlagen, in denen Kognitionen eine Ursache von Emotionen darstellen. Eine erste kognitive Theorie stellt die Schachter-Singer-Theorie der Emotionen dar (Schachter & Singer, 1962). Nach Schachter und Singer (1962) ist eine physiologische Reaktion nicht ausreichend, um eine Emotion zu verursachen. In ihrem Modell nahmen die Autoren an, dass ein Reiz eine physiologische Erregung auslöst und erst Kognitionen, die die physiologische Erregung erklären können, eine spezifische Emotion auslösen. Grundlage für diese Idee stellen experimentelle Untersuchungen dar, in denen Emotionen durch eine Manipulation von Kognitionen gezielt hervorgerufen werden konnten (Schachter & Singer, 1962).

### **2.7.2 Appraisal-Theorien**

Zeitlich parallel zur Schachter-Singer-Theorie wurde mit den sogenannten Appraisal-Theorien eine Gruppe weiterer kognitiver Emotionstheorien eingeführt. In Appraisal-Theorien wird angenommen, dass eine Emotion durch ein spezifisches Muster an kognitiven Bewertungen der Situation ausgelöst wird und nicht durch die Situation per se (Roseman, Spindel & Jose, 1990). Mit dieser Annahme konnten einige Probleme bisheriger

Emotionstheorien behandelt werden: Im Rahmen von Appraisal-Theorien ist es möglich, dass Personen unterschiedlich auf bestimmte Reize reagieren und dass sich die emotionalen Reaktionen eines Individuums mit der Zeit verändern können (Roseman & Smith, 2001). Zudem erlauben diese Theorien, dass verschiedene Situationen zur gleichen Emotion führen, nämlich dann, wenn die gleiche Bewertung zu Grunde liegt (Roseman & Smith, 2001). Anders als in der Schachter-Singer Theorie, in der angenommen wird, dass Kognitionen zur Emotionsauslösung bewusst sein müssen, können Bewertungen im Rahmen der Appraisal-Theorien sowohl kontrollierte als auch automatisch ablaufende Prozesse darstellen (Moors, 2009; Roseman & Smith, 2001)

Appraisal-Theorien stellen eine Gruppe von Theorien dar. Zwar ist den Theorien gemeinsam, dass Bewertungen von Situationen für die Auslösung einer Emotion ausschlaggebend sind, allerdings unterscheiden sich die Appraisal-Theorien auch hinsichtlich einiger wesentlicher Punkte (siehe z. B. Scherer, 1999). Vor allem unterscheiden sich die Theorien hinsichtlich der Anzahl und Art der angenommenen Bewertungsdimensionen.

Eine dieser Appraisal-Theorien ist die Theorie von Roseman (z. B. Roseman, 2001). Rosemans (2001) Appraisal Theorie spielt im Rahmen der Diagnostik von EU eine besondere Rolle. Anhand dieser Theorie haben MacCann und Roberts (2008) ein Verfahren zur Erfassung von EU entwickelt. In Abbildung 2.3 wird die Appraisal-Theorie von Roseman (2001) illustriert. In Rosemans Appraisal-Theorie werden 17 Emotionen (surprise, hope, fear, joy, relief, sadness, distress, frustration, disgust, dislike, anger, contempt, regret, guilt, shame, love, and pride) durch ein Muster von Bewertungen einer Situation auf sieben Appraisal-Dimensionen vorhergesagt (Roseman, 2001). Der Abbildung kann für jede der 17 Emotionen in den Zeilen und Spalten die Ausprägungen auf den Bewertungsdimensionen entnommen werden, die für die Auslösung angenommen werden. Nach Roseman (2001) bewerten Individuen Situationen hinsichtlich der folgenden sieben Dimensionen:

1. *Situational State*: Diese Bewertungsdimension bezieht sich darauf, ob eine Situation als konsistent oder inkonsistent mit den eigenen Motiven bewertet wird. Diese Dimension trennt zugleich auch zwischen positiven (*motive-consistent*) und negativen Emotionen (*motive-inconsistent*) (Roseman et al., 1990).
2. *Motivational State*: Auf dieser Dimension wird bewertet, ob das Motiv der Person in der Situation darin besteht, eine positive Konsequenz herbeizuführen (*appetitive*) oder eine negative Konsequenz zu vermeiden (*aversive*).
3. *Causal Agency*: Auf der Dimension Causal Agency wird abgebildet, wem die Verantwortung für die Situation zugeschrieben wird. Die Verursachung der Situation kann entweder einer anderen Person (*other*), der eigenen Person (*self*) oder den Umständen (*circumstances*) zugeschrieben werden.
4. *Expectedness*: Die Bewertung auf der Dimension Expectedness bezieht darauf, ob die gegebene Situation den Erwartungen entspricht oder nicht. Wie im Modell zu erkennen, trennt diese Dimension die Emotion Überraschung von allen anderen Emotionen.
5. *Certainty*: Certainty stellt eine Bewertung der Wahrscheinlichkeit des Eintretens einer Situation dar, mit den Ausprägungen des sicheren Eintretens (*certain*) oder unsicheren Eintretens (*uncertain*).
6. *Control Potential*: Auf dieser Dimension wird bewertet, ob die Situation kontrolliert werden kann oder nicht. Das Kontrollpotenzial kann entweder als hoch (*high*) oder gering (*low*) bewertet werden.
7. *Instrumental/ Intrinsic Problem*: Diese Bewertungsdimension bezieht sich nur auf Situationen, die als motiv-inkonsistent bewertet werden und damit negative Emotionen auslösen. Das Problem in der Situation kann entweder das Erreichen eines Ziels

blockieren (*instrumental*) oder aufgrund einer inhärenten Eigenschaften der Situation unerwünscht sein (*intrinsic*).

		1. Situational State					
		Motive-Consistent		Motive-Inconsistent			
		2. Motivational State					
3. Causal Agency	4. Expectedness/ 5. Certainty	Appetitive Motive	Aversive Motive	Appetitive Motive	Aversive Motive	6. Control Potential	
Circumstances	Unexpected	Surprise					
	Uncertain	Hope		Fear		Low	
	Certain	Joy	Relief	Sadness	Distress		
	Uncertain	Hope		Frustration	Disgust	High	
	Certain	Joy	Relief				
Other	Uncertain/ certain	Love		Dislike		Low	
	Uncertain/ certain			Anger	Contempt	High	
Self	Uncertain/ certain	Pride		Regret		Low	
	Uncertain/ certain			Guilt	Shame	High	
		Instrumental		Intrinsic			
		7. Problem Type					

Abbildung 2.3. Appraisal-Theorie von Roseman (2001).

Obschon Appraisal-Theorien weit akzeptiert sind, sind sie auch mit einigen Problemen behaftet. Zum einen fehlt den Theorien ein Konsens über die Art und Anzahl der Bewertungsdimensionen (Scherer, 1993). Bisher gibt es eine Reihe von Theorien, die sich teilweise deutlich unterscheiden (für eine Übersicht siehe Roseman et al., 1990; siehe auch Scherer, 1993). Zum anderen wurden die Methoden der Testung dieser Theorien in Frage gestellt, die auch auf die Weiterentwicklung der Theorien einen Einfluss haben (Scherer, 1999). Vor allem wurde in diesem Zusammenhang kritisiert, dass ein Großteil der Evidenz für die Theorien allein auf Selbstberichten von Personen basiert. In vielen Untersuchungen sollen sich Personen bspw. eine vergangene Situation vorstellen und diese anhand von Ratingskalen

bewerten, die die theoretisch angenommenen Bewertungsdimensionen widerspiegeln. Hierbei wurde kritisiert, dass Personen nicht immer einen Einblick in ihre Bewertungs-Prozesse haben, da sie teilweise automatisch ablaufen können oder die Erinnerung an die Bewertungen fehlerhaft sein kann (siehe Scherer, 1999). Allerdings konnte auch stärkere Evidenz für einige dieser Theorien aus experimentellen Untersuchungen verzeichnet werden (z. B. Roseman, 1991; Roseman & Evdokas, 2004). Experimentelle Untersuchungen sind oft so gestaltet, dass Personen Vignetten von Situationsbeschreibungen vorgelegt werden, die bezüglich ihrer Bewertungsdimensionen systematisch manipuliert wurden (z. B. Roseman, 1991; Scherer, 1999). Die Personen sollen dann vorhersagen, welche Emotionen aus dieser Situation resultieren und es kann daran überprüft werden, ob tatsächlich die aus der Theorie abgeleiteten Emotionen vorhergesagt werden. Insgesamt gibt es mittlerweile einige Evidenz dafür, dass kognitive Bewertungen Emotionen verursachen (Roseman & Evdokas, 2004), wobei die einzelnen Theorien in ihren Vorhersagen sehr unterschiedlich sind.

## **2.8 Die Messung von EI und EU**

Bei der Betrachtung von Operationalisierungen der EI kann derzeit ein Zustand beobachtet werden, der für eine Form von Intelligenz sehr ungewöhnlich ist: Ein großer Teil der Forschung konzentriert sich auf die Entwicklung von Selbstberichtsverfahren, während die Konstruktion von Leistungstests nur äußerst zögerlich voranschreitet. Mit der Operationalisierung von EI als Leistungstest oder Selbstberichtsverfahren haben sich zwei Richtungen herausgebildet, die nahezu unabhängig voneinander verfolgt werden. Diese beiden Richtungen sind mit der Art der Konzeptualisierung von EI assoziiert. Im Rahmen von Fähigkeitsmodellen werden Leistungstests entwickelt, während Modelle der EI als Mischung aus Persönlichkeitseigenschaften und Fähigkeiten in Form von Selbstberichten operationalisiert werden. Ein Grund für die hohe Anzahl verfügbarer Selbstberichtsverfahren



(siehe z. B. Pérez, Petrides & Furnham, 2005), die überdies weiterhin wächst, kann somit in konzeptuellen Unklarheiten im Bereich der EI gefunden werden.

Neben der Kritik an Selbstberichtsverfahren, die sich auf die damit zusammenhängende Art der Konzeptualisierung von EI als Persönlichkeitseigenschaft bezieht, wurde auch die grundsätzliche Angemessenheit der Verwendung von Selbstberichtsverfahren zur Erfassung einer Fähigkeit in Zweifel gezogen (z. B. Schulze et al., 2007). Die Verwendung von Selbstberichten der EI setzt voraus, dass Personen sowohl in der Lage als auch motiviert sind, ihre emotionsbezogenen Fähigkeiten akkurat einzuschätzen. Im Bereich der klassischen Intelligenz wurde festgestellt, dass Leistungstests und Selbsteinschätzungen allenfalls moderate, positive Zusammenhänge miteinander aufweisen (siehe Freund & Kasten, 2012). Auch hinsichtlich der EI gibt es Hinweise darauf, dass Selbstberichte die tatsächliche Fähigkeit der EI, erfasst mit einem Leistungstest, nicht akkurat abbilden (Sheldon, Dunning & Ames, 2014).

Im Bereich der Messung der EI als Fähigkeit kann dagegen ein Mangel an verfügbaren Instrumenten verzeichnet werden. Die leistungsbasierte Messung der EI wird derzeit durch einen einzelnen Ansatz dominiert: den MSCEIT und dessen Vorgänger, den MEIS (Mayer et al., 1999; Mayer et al., 2002). Beide Instrumente stellen jeweils Operationalisierungen aller vier im Four-Branch Modell der EI angenommenen Fähigkeitsbereiche dar. Darüber hinaus wurden noch einige weitere Leistungstests entwickelt, die allein zur Erfassung von EU dienen. Eines dieser Verfahren ist der STEU (MacCann & Roberts, 2008), der sich mittlerweile als ein recht vielversprechender Ansatz herauskristallisiert hat. Weitere Verfahren zur Erfassung von EU, die neben dem STEU vorgeschlagen wurden, wurden in den bisherigen Studien eher selten genannt oder eingesetzt.

Den meisten Instrumenten zur Erfassung von EU ist gemeinsam, dass sie als Situational Judgment Tests (SJTs) klassifiziert werden können. Ein wichtiges

Unterscheidungsmerkmal der bisherigen Verfahren liegt in der Art, wie mit der größten Herausforderung im Rahmen der Entwicklung von Leistungstests für EI umgegangen wurde: der Festlegung eines angemessenen Scoring-Kriteriums. Um die Leistung einer Testperson in einem Wert abbilden zu können, bedarf es eines Kriteriums, mit dem die Akkuratheit der Testantwort eindeutig feststellbar ist. Bei den meisten kognitiven Fähigkeitstests stellt die Identifikation eines veridikalen Scoring-Kriteriums kein größeres Problem dar. Im Zusammenhang mit emotionalen oder sozialen Fähigkeiten lässt sich ein solches Kriterium allerdings nur schwer identifizieren. Für Scoring-Kriterien wurden verschiedene Vorschläge gemacht (siehe Mayer et al., 1999), die allerdings mit Problemen behaftet sind und keineswegs zufriedenstellend erscheinen.

Angesichts des Mangels an Verfahren und geeigneten Scoring-Ansätzen ergeben sich für die Messung der EI mehrere Probleme. Im Sinne des Validitätskonzeptes von Messick (1995) ist der für die Konstruktvalidierung notwendige Zugang zu verschiedenartigen Quellen von Belegen erschwert. Dass zudem die bisherigen Erkenntnisse über EI und EU fast ausschließlich auf dem MSCEIT-/MEIS-Ansatz beruhen, ist ebenfalls kritisch zu bewerten, da Konstrukt und Methode somit untrennbar miteinander konfundiert sind. Glücklicherweise wurden in den letzten Jahren auch einige neuere Vorschläge gemacht, die den notwendigen Fortschritt im Bereich der EI und EU ermöglichen können.

### **2.8.1 Situational Judgment Tests**

Bei SJTs handelt es sich um eine Art von Messverfahren, in denen Situationsbeschreibungen präsentiert werden und anschließend die Reaktionen der Getesteten auf diese Situationen gemessen werden. Das charakteristische Merkmal von SJTs besteht darin, dass sie aus zwei Komponenten zusammengesetzt sind: einem sogenannten Item-Stamm, in dem die Situation präsentiert wird, sowie den Antwortoptionen (Weekley, Ployhart & Holtz, 2006). SJTs können zur Erfassung sehr unterschiedlicher Konzepte aus der Fähigkeits- und

Persönlichkeitsdomäne zum Einsatz kommen. Mit der Unterschiedlichkeit der intendierten Konzepte verbunden, können die Instruktionen in SJTs stark variieren. Bezüglich der Instruktionen wurden zwei Klassen unterschieden: das *Knowledge-* sowie das *Behavioral-Tendency-Format*. Beim Knowledge-Format (auch als *should-do-Instruktion* bezeichnet) wird die Testperson aufgefordert, eine Antwort zu geben, die von ihr als korrekt eingeschätzt wird (Weekley et al., 2006). Beispielsweise kann im Rahmen eines SJTs die Aufgabe darin bestehen, die effektivste Verhaltensweise hinsichtlich der Erreichung eines vorgegebenen Ziels auszuwählen. Dieses Antwortformat ist prinzipiell für die Erfassung maximalen Verhaltens geeignet. Beim Behavioral-Tendency-Format (auch als *would-do-Format* bezeichnet) soll die Testperson hingegen ihre typische Reaktion in der gegebenen Situation berichten, was zur Erfassung von Persönlichkeitseigenschaften dienen kann (Weekley et al., 2006). Darüber hinaus kann auch das Format der Antwortoptionen in SJTs variieren. Zumeist sollen die Testpersonen die vorgegebenen Optionen entweder anhand eines Forced-Choice-Antwortformates oder einer Likert-Skala bewerten (Weekley et al., 2006).

Obwohl SJTs bereits seit langer Zeit eingesetzt werden, ist das Interesse an dieser Methode insbesondere seit den 1990ern gestiegen (Weekley & Ployhart, 2006). SJTs werden häufig im Rahmen der berufsbezogenen Diagnostik genannt (z. B. McDaniel & Ngyuen, 2001), sind aber nicht auf diesen beschränkt (siehe z. B. Orchard et al., 2009). Das häufigste Anwendungsziel bei der Konstruktion von SJTs besteht in der Erfassung von interpersonalen Fähigkeiten (Christian, Edwards & Bradley, 2010). Im Bereich der EI kamen SJTs bisher vor allem zur Messung von EU und Emotional Management sowie vereinzelt für Perception zum Einsatz (Mayer et al., 1999; MacCann & Roberts, 2008). Zur Messung von EU werden den Testpersonen Situationsbeschreibungen vorgegeben, anhand derer sie korrekte Einschätzungen über die Auslösung bzw. den Verlauf von Emotionen oder über die Gründe von Emotionen treffen sollen.

Mit dem Begriff SJT wird oft auch ein bestimmtes Vorgehen im Rahmen der Konstruktion dieser Tests verbunden (z. B. McDaniel & Nguyen, 2001). Klassischerweise werden SJTs in drei Schritten konstruiert, wobei der Item-Stamm, die Antwortoptionen sowie der Antwortschlüssel in drei separaten Schritten von jeweils unterschiedlichen Gruppen von Expertinnen und Experten konstruiert werden (McDaniel & Nguyen, 2001). Da verschiedenste Alternativen zu diesem Vorgehen existieren (siehe Weekley et al., 2006), werden häufig sämtliche Tests als SJTs bezeichnet, die Situationsbeschreibungen als Testmaterial beinhalten (Orchard et al., 2009). Orchard et al. (2009) machten darauf aufmerksam, dass SJTs im Bereich der EI zumeist nicht dem klassischen Vorgehen entsprechend konstruiert wurden. Zudem fehlen sogar häufig Informationen über die Vorgehensweise bei der Konstruktion dieser Tests (siehe z. B. Mayer et al., 2002). Eine Ausnahme stellt hierbei der STEM von MacCann und Roberts (2008) dar. Der STEM ist ein SJT zur Erfassung von Emotional Management, der 44 Vignetten mit Situationsbeschreibungen sowie jeweils vier Handlungsoptionen umfasst, die hinsichtlich ihrer Effektivität bewertet werden sollen. Der Konstruktion von Item-Stämmen wurden semistrukturierte Interviews zugrunde gelegt, in denen zunächst eine erste Stichprobe von 55 Personen gebeten wurden, emotionale Situationen aus den letzten beiden Wochen zu berichten. Die daraus entwickelten Item-Stämme wurden danach einer zweiten Stichprobe von 99 Personen vorgelegt. Die Aufgabe der zweiten Stichprobe bestand darin, jeweils zu beschreiben, a) welche Handlungsweise sie für die beste in der gegebenen Situation halten und b) wie sie sich selbst in der Situation verhalten würden. Durch die Verwendung von Paradigmen der SJT-Forschung konnte der STEM systematisch erstellt werden und beinhaltet realitätsnahe Situationen. Das Problem der Identifikation eines geeigneten Antwortschlüssels bleibt allerdings auch bei Tests, wie dem STEM, die unter Verwendung von SJT-Paradigmen konstruiert werden, bestehen.

## **2.8.2 Scoring**

Aufgrund des Mangels an einem veridikalen Kriterium für die Bewertung von Testantworten, stellt das Scoring von Tests der EI eine besondere Herausforderung dar. Um das fehlende veridikale Scoring-Kriterium zu ersetzen, wurden einige alternative Bewertungsstandards vorgeschlagen. Als Scoring-Kriterien für Tests der EI wurden Urteile von Expertinnen und Experten, Urteile großer Gruppen von Novizen, Selbstberichte von Zielpersonen sowie, als relativ neuer Ansatz, Emotionstheorien herangezogen.

### ***2.8.2.1 Experten-Scoring***

Beim Experten-Scoring stellen Urteile von Expertinnen und Experten das Kriterium für die Richtigkeit der Testantworten dar. Eine Antwort wird dann als korrekt bewertet, wenn sie mit dem Mehrheitsurteil von Expertinnen und Experten übereinstimmt. Ein großes Problem dieses Ansatzes besteht darin, Expertinnen und Experten für den Bereich der Emotionen zu identifizieren. Das Experten-Scoring ist bisher u. a. beim MEIS und MSCEIT zum Einsatz gekommen. Im MEIS lieferten zwei der Testautoren und im MSCEIT 21 Emotionsforscher der International Society of Research in Emotions die Expertenurteile für das Scoring. Anhand beider Tests lässt sich das Problem der Auswahl einer Experten-Gruppe veranschaulichen. Die hinter der Entscheidung für die jeweilige Experten-Gruppe stehende Annahme, dass Forscherinnen und Forscher im Bereich der EI oder der Emotionen tatsächlich korrekte Antworten in einem Test der EI liefern, erscheint zweifelhaft. Matthews et al. (2002) wiesen darauf hin, dass die Forschungsgrundlage im Bereich der Emotionen, auf deren Basis die genannten Expertinnen und Experten ihre Urteile treffen könnten, nur begrenzt ist und eher allgemeingültige Aussagen auf Grundlage statistischer Zusammenhänge umfasst. Für die Beantwortung von Items, die sich bspw. auf die effektivste Möglichkeit der Emotionsregulation für ein bestimmtes Individuum in einer konkreten Situation bezieht, erscheint diese Grundlage nicht ausreichend zu sein. Expertinnen und Experten sind sich in

diesen Bereichen zudem in vielen Punkten uneinig (Barchard & Russell, 2006; Legree, Psozka, Tremble & Bourne, 2005), was die Zweifel daran verstärkt, dass die akkurate Antwort mit diesem Kriterium eindeutig identifiziert werden kann. Zwar ist es möglich, dass es bestimmte (Berufs-)Gruppen gibt, die eine höhere Expertise im Bereich der Emotionen aufweisen; ob die Identifizierung einer Expertengruppe letztendlich erfolgreich war, kann bisher allerdings nicht überprüft werden, angesichts der Tatsache, dass es an einem objektiven Standard fehlt (Barchard, Hensley & Anderson, 2013).

### ***2.8.2.2. Consensus-Scoring***

Für Domänen, in denen sowohl ein objektives Kriterium der Akkuratheit als auch Expertinnen und Experten schwer zu identifizieren sind, wie es für den Bereich der EI der Fall ist, wurde das sogenannte Consensus-Scoring vorgeschlagen (Lagree et al., 2005). Beim Consensus-Scoring wird die Antwort einer Testperson anhand der Antwortverteilung einer großen Gruppe von Novizen bewertet. Diese Normgruppe kann aus der aktuell getesteten Stichprobe selbst oder einer hiervon unabhängigen Normstichprobe bestehen. Als Kriterium für die Generierung des Antwortschlüssels werden die Häufigkeitsverteilungen der Item-Antworten der Novizen herangezogen. Dabei gibt es unterschiedliche Arten, wie die Testantwort einer Person anhand der Antwortverteilung der Normgruppe mit Punkten bewertet werden kann. Beim sogenannten Proportional Consensus Scoring wird der Testantwort einer Person ein Wert zugewiesen, der proportional zur Häufigkeit dieser Antwort in der Bezugs-Stichprobe ist. Beispielsweise könnte eine Person beim einem Item den Wert .69 erhalten, wenn 69% der Stichprobe ebenfalls diese Testantwort gegeben haben, während sie für eine andere Option, die nur von 20% der Normstichprobe gewählt wurde, einen Wert von .20 erhielt. Beim Mode-Scoring wird hingegen allein die am häufigsten gewählte Antwort als korrekt und mit einem Punkt bewertet, während alle anderen Optionen null Punkte erhalten (im Fall einer Ratingskala werden in einer Variante dieses Verfahrens auch die zur häufigsten Antwort

benachbarten Skalenpunkte als korrekt bewertet). Bei Rating-Skalen wurde zudem auch das Distanz-Scoring eingesetzt, bei dem die Leistung in einem Item über die Nähe der Testantwort zur häufigsten Antwort abgebildet wird (für eine Übersicht verschiedener Scoring-Varianten siehe Barchard et al., 2013; siehe auch MacCann et al., 2004).

Das Consensus-Scoring wurde bisher in Testverfahren für verschiedene Domänen eingesetzt, wie zur Messung von EI, sozialer Intelligenz, Führungsfähigkeit oder Wissen über bestimmte Aspekte von Fahrsicherheit (siehe Legree et al., 2005). Laut Legree et al. (2005) beinhalten Domänen, für die das Consensus Scoring besonders geeignet ist, vor allem implizite Wissensinhalte, die eher durch Erfahrung im Alltag, als durch formale Edukation vermittelt werden. Im Bereich der EI kam das Consensus-Scoring bspw. im MEIS und MSCEIT zum Einsatz. Dem Consensus-Scoring liegt die Annahme zu Grunde, dass Urteile von Novizen- und Experten-Gruppen hinsichtlich ihrer zentralen Tendenz übereinstimmen und sich nur insofern voneinander unterscheiden, dass die Antworten von Novizen mehr streuen als die Urteile einer Expertengruppe (Legree et al., 2005). Die zentrale Tendenz der Urteile von Experten- und Novizen-Gruppe soll dabei mit der korrekten Antwort übereinstimmen (Legree et al., 2005).

Als empirischen Beleg für diese Annahme wurde zumeist entweder die Übereinstimmung der Antwortschlüssel nach Experten- oder Consensus-Scoring oder die Korrelation zwischen den Testwerten, einmal über Consensus-Scoring und einmal über Experten-Scoring ausgewertet, herangezogen. Tatsächlich konnten für verschiedene Tests sehr hohe Korrelation zwischen den nach Consensus- und Experten-Scoring ausgewerteten Testantworten festgestellt werden (Mayer et al., 2003). Im MSCEIT bspw. lagen die Korrelationen zwischen Experten- und Consensus-Scores für die Unterskalen sowie den Gesamtwert zwischen  $r = .96$  und  $.98$ , woraus die Schlussfolgerung gezogen wurde, dass Consensus- und Experten-Scoring austauschbare Alternativen seien (Mayer et al., 2003). Den

Erwartungen entsprechend waren zudem die Übereinstimmungen der Urteile innerhalb der Experten-Gruppe höher, als die Übereinstimmungen innerhalb der Novizen-Gruppe (Legree et al., 2005; Mayer et al., 2003).

Die im MSCEIT beobachtete hohe Korrelation zwischen den nach Consensus- und Experten-Scoring ermittelten Testwerten ist allerdings kein ausreichender Beleg dafür, dass das Consensus-Scoring geeignet ist, die korrekte Antwort zu identifizieren. Zum einen gibt es trotz einiger Belege für eine hohe Korrelation beim MSCEIT auch Hinweise darauf, dass ein solcher Zusammenhang nicht immer zu erwarten ist. Für den MEIS wurden deutlich geringere Korrelationen zwischen den auf Experten- und Consensus-Scoring basierten Testwerten gefunden (Roberts et al., 2001). Zum anderen ist es möglich, dass weder die Mehrheitsurteile der Novizen noch die der Expertinnen und Experten mit der korrekten Antwort übereinstimmen. Eine hohe Übereinstimmung zwischen Urteilen von Experten- und Novizen ist auch dann zu erwarten, wenn Expertinnen und Experten nicht erfolgreich identifiziert werden konnten und die Experten-Gruppe sich hinsichtlich ihrer Fähigkeit nicht von den Novizen unterscheidet. Zudem ist es möglich, dass sowohl die Übereinstimmung zwischen Experten- und Consensus-Scoring als auch die Güte der beiden Scoring-Ansätze von Eigenschaften der jeweiligen Tests abhängen, wie bspw. den Schwierigkeiten der Items (Barchard et al., 2013).

In anderen Studien wurde das Consensus-Scoring mit Tests durchgeführt, die gleichzeitig auch ein veridikales Scoring-Kriterium bieten, so dass die Qualität des Consensus-Scorings anhand eines eindeutigen Standards beurteilt werden kann. Mohorić, Takšić und Duran (2010) sowie Barchard et al. (2013) haben hierzu Vocabulary-Tests herangezogen, für die die Sprache ein veridikales Regelwerk zur Beurteilung der Testantworten liefert. Mohorić et al. (2010) führten einen Vergleich verschiedener Scoring-Varianten mit einem Vocabulary-Test für Emotionsbegriffe durch, den Vocabulary Emotions



Test (VET), der von den Autorinnen und Autoren als ein neues Maß für EU vorgeschlagen wurde. Für den VET ergaben sich erste Hinweise darauf, dass die Testwerte, die über Varianten des Consensus-Scorings ermittelt wurden, teilweise nur moderate positive Beziehungen mit den Testwerten auf Basis des veridikalen Standards, einem kroatischen Wörterbuch, aufwiesen (Mohorić et al., 2010). In einer Studie von Barchard et al. (2013) wurde ein etablierter, allgemeiner Vocabulary-Test einmal anhand des veridikalen Kriteriums und einmal anhand des Consensus-Scorings ausgewertet. Diese Studie erbrachte Hinweise darauf, dass die Übereinstimmung des veridikalen Scorings mit dem Consensus-Scoring mit steigender Schwierigkeit der Items abnimmt. Während die Korrelation der Testwerte nach dem Consensus Scoring mit den Testwerten auf Basis des veridikalen Scoring für ein Set leichter Items nahe 1 lag, fiel diese für eine Gruppe relativ schwerer Items deutlich geringer aus ( $r = .53$ ). Barchard et al. (2013) schlugen anhand der Ergebnisse ihrer Studie vor, dass ein Test, bei dem auf das Consensus-Scoring zurückgegriffen werden soll, aus vielen leichten Items zusammengesetzt sein sollte. Da allerdings die tatsächlichen Item-Schwierigkeiten nur dann ermittelt werden können, wenn ein veridikales Scoring-Kriterium vorliegt, kann diese Empfehlung im Bereich der EI nur schwer umgesetzt werden.

Darüber hinaus machten MacCann et al. (2004) auf ein Problem des Consensus Scorings aufmerksam, das sich auf die Form der Verteilung der Testwerte bezieht. In der Studie von MacCann et al. (2004) wurde das Ergebnis erbracht, dass unter Verwendung des Consensus-Scorings nur schlecht zwischen Personen differenziert werden, die sehr hohe Fähigkeiten aufweisen (MacCann et al., 2004). Barchard und Russell (2006) zogen zudem die Fairness von Tests unter Verwendung des Consensus-Scorings in Zweifel. Sie argumentierten, dass das Consensus-Scoring gegenüber Mitgliedern von Untergruppen, die innerhalb der Normstichprobe eine Minderheit darstellen, einen Bias aufweisen kann, vor allem unter Verwendung des Mode-Scorings. Unterscheiden sich Subgruppen hinsichtlich

ihrer häufigsten Antwort voneinander, hat unter dem Mode-Scoring die Gruppengröße einen Einfluss auf die mittleren Testwerte der Untergruppen (Barchard & Russell, 2006).

Insgesamt erscheint es derzeit fraglich, ob das Consensus-Scoring eine sinnvolle Scoring-Alternative bieten kann und falls ja, unter welchen Bedingungen dies möglich ist. Auch Befürworter dieses Verfahren weisen darauf hin, dass zukünftige Studien klären müssen, ob die im Consensus-Scoring getroffenen Annahmen tatsächlich zutreffend sind (siehe Legree et al., 2005). Die bisherigen Studien lassen zumindest begründete Zweifel daran aufkommen.

### ***2.8.2.3 Target-Scoring***

Das Target-Scoring wurde schon sehr früh als eine Möglichkeit zur Auswertung von Tests sozio-emotionaler Fähigkeiten vorgeschlagen (siehe Cronbach, 1955). Beim Target-Scoring stellt die Auskunft einer Zielperson über ihre eigenen Emotionen das Kriterium für die Bewertung von Testantworten dar. Die Annahme beim Target-Scoring ist, dass eine Zielperson direkter, und damit besser Einsicht in ihre eigenen Emotionen nehmen kann, als außenstehende Personen, die die Emotionen der Zielperson einschätzen sollen.

Außerhalb des Bereichs der EI wurde das Target-Scoring zur Messung von Empathie-Fähigkeit und sozialer Intelligenz eingesetzt (Ickes, 2001; Weis & Süß, 2005). Bei der Messung von EI wurde das Target-Scoring bisher nur für einige Unterskalen zur Messung von Perception im MEIS sowie für einen Test zur Messung von EU, dem Test of Emotional Intelligence (TEMINT; Blickle, Momm, Kramer, Mierke, Liu & Ferris, 2009), verwendet.

Ein Problem beim Target-Scoring besteht darin, dass die Zielperson nicht in der Lage und/oder motiviert sein könnte, ihre Emotionen akkurat zu berichten. Eine Studie von Mayer und Geher (1996) erbrachte Belege dafür, dass Emotionen, die von Zielpersonen über sich selbst berichten werden, sozial erwünschter sind, als die Emotionen, die außenstehende

Personen bei der Zielperson einschätzen. Dies könnte ein Hinweis darauf sein, dass bei den Zielpersonen eine Verzerrung in Richtung sozialer Erwünschtheit vorliegt, obschon das Ergebnis prinzipiell auch durch Verzerrungen in den Fremdeinschätzungen erklärt werden könnte. In Anbetracht der Tatsache, dass a) Selbstberichte unter Verdacht von Antwortverfälschungen stehen und dass b) die korrekte Wahrnehmung der eigenen Emotionen ein Teil eines Fähigkeits-Konstruktes der EI selbst darstellt, das bei verschiedenen Zielpersonen unterschiedlich ausgeprägt sein kann, erscheinen Zweifel gegenüber dem Target-Scoring allerdings als berechtigt.

#### ***2.8.2.4 Theoriebasiertes Scoring***

Im Zusammenhang mit der Entwicklung des STEU wurde von MacCann und Roberts (2008) das theoriebasierte Scoring vorgeschlagen. Im Rahmen dieses Scoring-Ansatzes werden Emotionstheorien als Kriterium für die Bewertung von Testantworten in Tests für EI herangezogen. Zur Messung von EU wurden hierbei Theorien über die Auslösung von Emotionen als relevant erachtet. MacCann und Roberts (2008) zogen beim STEU eine bestimmte Appraisal-Theorie, nämlich Rosemans Appraisal-Theorie, als Grundlage für die Konstruktion von Items und das Scoring heran.

Ein Problem des theorienbasierten Scoring ist in dem Umstand zu finden, dass es sehr viele Emotionstheorien gibt (unter diesen auch viele Appraisal-Theorien), in denen teilweise sich widersprechende Aussagen über Emotionen und deren Bedingungen zur Auslösung getroffen werden. Die Vielfalt der Emotionstheorien ist auch ein Grund dafür, dass die Grundlage für Expertinnen und Experten beim Experten-Scoring als unsicher bezeichnet werden kann. Trotz dieser Probleme ist das theoriebasierte Scoring bisher der Ansatz, der am besten begründet erscheint. Für eine Weiterentwicklung dieses Ansatzes wäre es interessant, Konvergenzen zwischen den bisherigen Appraisal-Theorien zu suchen und auf deren Basis einen Test entwickelt.

### **2.8.3 Leistungstests zur Erfassung von EI und EU**

In Leistungstests der EI werden die Fähigkeiten der Testpersonen gemessen, indem emotionsbezogene Aufgaben vorgelegt werden, die von den Testpersonen korrekt gelöst werden müssen. Da ein Item in einem Leistungstest das maximale Verhalten der Testperson provoziert, erlaubt diese Art der Messung, im Gegensatz zu Selbstberichten, eine direkte Beobachtung der Fähigkeit. In den folgenden Abschnitten werden die aktuell verfügbaren Leistungstests für EI und EU näher vorgestellt. In Tabelle 2.1 wird zudem eine Übersicht der Tests bzw. Untertests für EU präsentiert, zusammen mit Beschreibungen des Aufgabenmaterials sowie des Scorings.

Tabelle 2.1

*Übersicht der Messinstrumente zur Erfassung von EU*

Test	Untertest	Aufgabe	Items	Scoring
MSCEIT	Blends	Entweder wird a) eine komplexe Emotion oder b) Bestandteile einer komplexen Emotion verbal vorgegeben. Danach sollen entweder a) die passenden Bestandteile der komplexen Emotionen oder b) die komplexe Emotion aus einer Reihe von Optionen gewählt werden.	12 Multiple-Choice-Items	Consensus oder Experten
MSCEIT	Changes	Emotionen werden verbal vorgegeben und die Testperson soll eine Emotion aus einer Reihe von Emotionen wählen, die durch Zunahme der Intensität der vorgegebenen Emotionen resultieren könnte.	20 Multiple-Choice-Items	Consensus oder Experten
MEIS	Blends	Eine komplexe Emotion wird vorgegeben. Die Bestandteile dieser Emotion soll aus vier Optionen gewählt werden. Beispiel: „Optimism most closely combines which two	8 Multiple-Choice-Items	Consensus oder Experten

Test	Untertest	Aufgabe	Items	Scoring
		emotions?  (a) pleasure and anticipation  (b) acceptance and joy  (c) surprise and joy  (d) pleasure and joy.“ (Mayer et al., 1999, S. 276)		
MEIS	Progression	Ein Verlauf von Emotionen wird vorgegeben. Dazu soll eine Emotion aus vier Optionen gewählt werden, die auf den Verlauf der Emotionen folgt.  Beispiel: „If you feel angrier and angrier toward someone so that you are losing control, it would result in (choose one):  (a) gloating  (b) resentment  (c) hate  (d) rage.“  (Mayer et al., 1999, S. 276)	8 Multiple-Choice-Items	Consensus oder Experten
MEIS	Transitions	Es wird eine Emotion vorgegeben, die zu Beginn vorlag sowie eine weitere, die am Ende vorlag. Auf einer 5-	4 Situationen mit jeweils 6 Ratings	Consensus oder Experten

Test	Untertest	Aufgabe	Items	Scoring
		<p>Punkte Skala soll die Testperson die Wahrscheinlichkeiten einer Reihe von Emotionen, die zwischenzeitlich aufgetreten sein könnten, einschätzen.</p> <p>Beispiel: „A person is afraid and later is calm. In between, what are the likely ways the person might feel?“</p> <p>(Mayer et al., 1999, S. 277)</p>		
MEIS	Relativity	<p>Eine Begegnung zwischen zwei Personen wird beschrieben. Auf einer 5-Punkte-Skala sollen jeweils für eine Reihe von Emotionen die Wahrscheinlichkeiten eingeschätzt werden, mit der diese bei den Zielpersonen aufgetreten sind.</p> <p>Beispiel: „A dog is chasing sticks outside when he runs out in the street and gets hit by a car. The driver stops when the dog's owner dashes over to check on</p>	<p>4 Situationen mit jeweils 10 Ratings</p>	<p>Consensus oder Experten</p>

Test	Untertest	Aufgabe	Items	Scoring
		<p>the dog.“</p> <p>Beispiel für eine mögliche Reaktion von einer der beiden Zielpersonen:</p> <p>„ashamed about not being able to have better trained the dog“ (Mayer et al., 1999, S. 277)</p>		
STEU	-	<p>Den Testpersonen werden kurze Situationsbeschreibungen vorgelegt. Es soll diejenige Emotion aus 5 möglichen ausgewählt werden, die die Zielperson am wahrscheinlichsten empfindet.</p> <p>Beispiel: „An irritating neighbor of Eve’s moves to another state. Eve is most likely to feel? (a) regret, (b) hope, (c) relief, (d) sadness, (e) joy“ (MacCann &amp; Roberts, 2008, S. 542)</p>	42 Multiple-Choice-Items	Theorie
TEMINT	-	<p>Den Testpersonen werden kurze Situationsbeschreibungen vorgelegt, die aus Sicht einer Zielperson berichtet werden.</p>	12 Situationen mit 10 Ratings	Target



Test	Untertest	Aufgabe	Items	Scoring
		<p>Für die Zielperson sollen danach die Intensitäten von 10 Emotionen, die möglicherweise resultieren, auf einer 3-Punkte-Skala eingeschätzt werden.</p> <p>„A 30-year-old female computer specialist reports: ‚My cat was ill. I had to take him to the surgeon. I thought I had poisoned him with insect spray.‘ How did this person feel in this situation?“</p> <p>Einzuschätzende Emotionen: „dislike, anger, fear, unease, sadness, guilt, happiness, pride, affection, and surprise.“ (Blickle et al., 2009, S. 103)</p>		
ACES/	Emotion	Kurze Vignetten mit	15 Multiple-	Festlegung
EKT	Situation	<p>Situationsbeschreibungen einer Zielperson werden vorgegeben, wie bspw. eine Situation, in der einem Kind ein Eis herunterfällt.</p> <p>Das getestete Kind soll eine von</p>	Choice Items	durch Testautoren

Test	Untertest	Aufgabe	Items	Scoring
		4 emotionalen Zuständen wählen, die von der Zielperson erlebt werden könnte.		

*Anmerkungen:* MSCEIT = Mayer-Salovey-Caruso Emotional Intelligence Test; MEIS = Multifactor Emotional Intelligence Scale; STEU = Situation Test of Emotional Understanding; TEMINT = Test of Emotional Intelligence; ACES = Assessment of Children's Emotion Skills; EKT = Emotional Knowledge Test. In der Spalte „Scoring“ werden die Scoring-Methoden berichtet, die von den Testautorinnen und -autoren vorgeschlagen wurden.

### **2.8.3.1 MSCEIT und MEIS**

Sowohl der MSCEIT als auch der MEIS stellen jeweils Versuche dar, das Four-Branch Modell der EI zu operationalisieren. Mit dem MEIS wurde von Mayer et al. im Jahre 1999 der erste leistungsbasierte Test zur Erfassung von EI zur Verfügung gestellt. Dieser diente wenige Jahre später als Grundlage für die Entwicklung des MSCEIT (Mayer et al., 2002). Der MSCEIT kann aktuell als das Standardverfahren zur Messung von EI bezeichnet werden. Während der MEIS als graue Literatur nur über Anfrage an die Testautoren verfügbar ist, ist der MSCEIT über einem Verlag veröffentlicht. Als ein Hauptziel bei der Entwicklung des MSCEIT wurde die Reduzierung des Umfangs des Vorgängers genannt (Mayer et al., 2002). Der MSCEIT ist mit 8 Untertests, die insgesamt 141 Items umfassen, deutlich kürzer als der MEIS, der 402 Items, verteilt auf 12 Aufgabentypen, enthält. Zudem erschien die psychometrische Qualität des MEIS in einigen Studien als unbefriedigend (siehe Roberts et al., 2001), was außerdem ein wichtiger Grund für eine Weiterentwicklung gewesen sein mag.

Im aktuellen Verfahren, dem MSCEIT, werden die vier Branches durch jeweils zwei Aufgabengruppen erfasst. Der angenommenen Struktur des Four-Branch Modells entsprechend, lassen sich Testwerte auf mehreren Ebenen ermitteln: ein EI-Gesamtwert, zwei Testwerte für die Bereiche Experiential- und Strategic-EI sowie vier Testwerte auf Ebene der Branches (Mayer et al., 2002). Zudem wird auch eine Rückmeldung auf Ebene der acht Aufgabengruppen angeboten, wobei die Autoren darauf hinweisen, dass diese Testwerte aufgrund der geringeren Messgenauigkeit mit Vorsicht zu interpretieren sind (Mayer et al., 2002). Zusätzlich stellt sich bei einer Rückmeldung auf Ebene der Aufgabengruppen die Frage, wie die Testwerte hier inhaltlich zu interpretieren sind, angesichts der Tatsache, dass wenig über die Struktur innerhalb der Branches bekannt ist.

Für einen Leistungstest weist das Testmaterial des MSCEIT (sowie auch das des MEIS) ein ungewöhnliches Merkmal auf: Die meisten Items verlangen ein Urteil auf einer Rating-Skala. Allein die beiden Subtests zur Erfassung von EU (bzw. zwei von vier Untertests zur Erfassung von EU im MEIS) haben ein Multiple-Choice Antwortformat. MacCann und Roberts (2008) machten auf das Problem aufmerksam, dass unter diesem Umstand der Inhalt des Branch mit dem Antwortformat konfundiert ist. Sie lieferten zudem Belege dafür, dass eine Veränderung des Antwortformats (Multiple-Choice vs. Rating-Skala) die korrelativen Beziehungen eines Tests für EI zu klassischer Intelligenz verändern kann. Diese Beobachtung ist inhaltlich derzeit schwer zu interpretieren, verdeutlicht aber, dass durch eine Konfundierung von Branch und Antwortformat Interpretationsprobleme entstehen können.

Perception wird im MSCEIT über Items erfasst, in denen in Gesichtern (Untertest *Faces*) sowie in landschaftlichen oder abstrakten Abbildungen (Untertest *Designs*) ausgedrückte Emotionen auf einer Ratingskala eingeschätzt werden sollen. Facilitation wird mit den Untertests *Facilitation* und *Sensations* gemessen. In der Aufgabengruppe Facilitation soll eingeschätzt werden, wie hilfreich Stimmungen für bestimmte Arten von kognitiven

Aufgaben sind. Im Untertest Sensations soll sich die Testperson eine Emotion vorstellen und diese mit anderen Wahrnehmungen, wie Farben oder Temperaturen, vergleichen.

EU wird im MSCEIT mit den Untertests *Blends* und *Changes* gemessen. Diese beiden Aufgabengruppen zielen auf zwei Aspekte ab, die im Four-Branch Modell unter den repräsentativen Fähigkeiten genannt werden (Mayer & Salovey, 1997): das Verstehen der Komplexität von Emotionen (*Blends*) und das Verstehen der zeitlichen Entwicklung von Emotionen (*Changes*). In Tabelle 2.1. wird das Testmaterial für die MSCEIT-Untertests *Blends* und *Changes* näher beschrieben. Zusätzlich können in dieser Tabelle auch Beschreibungen und Item-Beispiele der EU-Untertests des MEIS entnommen werden. Da es sich beim MSCEIT um geschütztes Material handelt, können hier keine Item-Beispiele präsentiert werden. Allerdings kann das Item-Material des MSCEIT anhand ähnlicher Aufgaben im MEIS veranschaulicht werden (siehe Mayer et al., 1999). Der Untertest *Blends* ist in MSCEIT und MEIS vergleichbar gestaltet. Der MSCEIT-Untertest *Changes* umfasst kurze Situationsbeschreibungen und entspricht hinsichtlich seiner Aufgabenstellung in etwa den MEIS-Untertests *Progression* sowie *Transitions* (siehe Mayer et al., 1999; Mayer et al., 2002).

Zur Messung von Management dienen im MSCEIT die beiden Aufgabengruppen *Management* und *Relations*, in denen den Testpersonen jeweils Situationsbeschreibungen mit mehreren Handlungsoptionen vorgelegt werden. Für jede Handlungsoption soll die Effektivität hinsichtlich der Emotionsregulation in der gegebenen Situation eingeschätzt werden. Im Untertest *Management* bezieht sich die Fragestellung dabei auf die eigene Person, während es im Untertest *Relations* um die Emotionsregulation einer fiktiven Zielperson geht.

Für die Auswertung des MSCEIT stehen zwei Optionen zur Verfügung: das Experten-Scoring anhand der Urteile von 21 Emotionsforschern oder das Consensus-Scoring, dem im MSCEIT eine Stichprobe von 5000 Personen zugrunde liegt. Die Auswertung der

Testantworten kann dabei nicht von den Anwenderinnen und Anwendern selbst vorgenommen werden, sondern wird über den Testverlag durchgeführt. Durch dieses Vorgehen ist der Antwortschlüssel verborgen, was die Forschung zum MSCEIT erschwert, da so der Zugang zu einigen Informationen, wie bspw. Item-Statistiken, erschwert ist.

Bezüglich der Reliabilität des MSCEIT wurden im Manual des Tests zufriedenstellende Ergebnisse berichtet (Mayer et al., 2002). Unter Verwendung des Consensus-Scorings wurden Split-Half-Reliabilitätskoeffizienten von .93 für den Gesamt-Score sowie .79 bis .91 auf Ebene der Branches erbracht. Vergleichbare Ergebnisse wurden unter dem Experten-Scoring erzielt (Mayer et al., 2002). Auf Ebene der Aufgabengruppen hingegen weist der MSCEIT eine deutlich geringe Messgenauigkeit auf, was von einigen Autorinnen und Autoren kritisiert wurde (Keele & Bell, 2009; Mayer et al., 2002). Ähnliche Ergebnisse hinsichtlich der Reliabilität wurden auch für eine deutsche Adaptation des MSCEIT erbracht (Steinmayr, Schütz, Hertel & Schröder-Abé, 2011). Da im Manual der deutschsprachigen Adaptation nur wenige Belege für die Validität berichtet werden und diese sich hauptsächlich auf die englischsprachige Version beziehen, wird die deutschsprachige Adaptation nicht näher betrachtet.

In Bezug auf die faktorielle Struktur des Tests, liegen heterogene Befunde vor. Wie bereits in Kapitel 2.1.3 (Struktur des Four-Branch Modells) erwähnt, konnte die angenommene Faktorenstruktur in einigen Studien nicht bestätigt werden, wobei vor allem die Belege für den Faktor Facilitation als ungenügend bewertet werden müssen.

Der MSCEIT weist moderate positive Beziehungen mit klassischer Intelligenz auf, die für einen Tests erwartet werden, der eine neue Form von Intelligenz messen soll. Hierbei kann über verschiedene Studien hinweg eine besondere Nähe zwischen EU und Gc festgestellt werden (Roberts et al., 2008). Darüber hinaus sind die Branches im MSCEIT weitestgehend unkorreliert mit den Big-Five Persönlichkeitsfaktoren, mit Ausnahme von Offenheit für

Erfahrung, was als diskriminanter Beleg für die Validität des MSCEIT gewertet werden kann (Roberts et al., 2008).

Insgesamt weist der MSCEIT, neben dem gravierenden Problem des Scorings, noch einige Schwächen hinsichtlich seiner Reliabilität und seiner faktoriellen Struktur auf. Zudem wirken einige der Untertests, wie Sensations oder Designs bezüglich ihres Inhaltes aus theoretischer Perspektive als wenig begründet. Bezüglich der Messung von EU im MSCEIT sind die beiden Untertest inhaltlich auf zwei der im Four-Branch Modell beschriebenen Fähigkeiten begrenzt, die zudem inhaltlich sehr unterschiedlich erscheinen.

Angesichts der genannten Probleme und des bisher eher geringen Erkenntnisstands im Bereich der EI, erscheinen die im Manual beschriebenen Anwendungsmöglichkeiten des MSCEIT, wie bspw. als zusätzliches Instrument bei der Personalauswahl, als deutlich verfrüht. Trotz dieser Schwächen wurde mit dem MSCEIT/MEIS-Ansatz ein wichtiger Beitrag für die Forschung zur EI geleistet, da mit diesem die erste leistungsorientierte und bis heute umfassendste Messung der EI ermöglicht wurde.

### **2.8.3.2 STEU**

MacCann und Roberts (2008) schlugen mit dem STEU einen theoriebasierten Ansatz zur Messung von EU vor. Der STEU ist ein SJT, der 42 Items umfasst, in denen zum einen Situationen aus dem persönlichen bzw. berufsbezogenen Alltag und zum anderen eher abstrakte Situationen beschrieben werden. Mittlerweile wurde auch eine Kurzversion des STEU, der STEU-brief, mit 19 Items vorgeschlagen (Allen, Weissman, Hellwig, MacCann & Roberts, 2014). Die SJT-Items werden von den Testpersonen in einem Multiple-Choice-Antwortformat beantwortet. In Tabelle 2.1 wird ein Item-Beispiel für den STEU präsentiert. Sowohl für die Konstruktion der Items als auch für das Scoring lieferte Rosemans Appraisal-Theorie die Grundlage. Eine Testantwort wird dann als korrekt bewertet, wenn diese mit den Aussagen von Roseman Appraisal-Theorie über die Zusammenhänge von

Ausprägungsmustern auf den Bewertungsdimensionen und den ausgelösten Emotionen übereinstimmt.

Hinsichtlich der Reliabilität des STEU berichteten MacCann und Roberts (2008) eine Schätzung für Cronbachs  $\alpha$  von .71. In einigen Fällen wurden allerdings auch deutlich geringere Werte ermittelt, wie in einer Studie von Austin (2010), in der die Schätzung bei  $\alpha = .48$  lag (siehe auch Studie 2 in MacCann & Roberts, 2008). Für ein Forschungsinstrument lagen die Reliabilitätsschätzungen in den meisten Studien allerdings in einem akzeptablen Bereich (Ferguson & Austin, 2010; MacCann, Pearce & Roberts, 2011; Libbrecht & Lievens, 2012).

Für den STEU wurden bereits einige konvergente und diskriminante Validitätsbelege erbracht. Mit anderen Maßen für EU wurden erwartungskonforme, positive Zusammenhänge gefunden: Der STEU ist mit der Branch EU aus dem MSCEIT zu  $r = .44$  korreliert (Austin, 2010). Für einen Subtest des MEIS (Stories) wurde ein Zusammenhang von  $r_{ab.c} = .33$  gefunden, nachdem sowohl Gc als auch die Big-Five Persönlichkeitsfaktoren statistisch kontrolliert wurden (MacCann & Roberts, 2008). Auch hinsichtlich der Beziehungen des STEU zu klassischer Intelligenz und den Big-Five wurden Ergebnisse beobachtet, die für ein Maß für EU zu erwarten sind. Zwischen dem STEU und Maßen für klassische Intelligenz wurden kleine bis moderate positive Zusammenhänge gefunden. Dabei erscheint der STEU, ähnlich wie der Branch EU im MSCEIT, eher mit Gc als mit Gf korreliert zu sein (Austin, 2010), wobei das Bild hier heterogen erscheint. Mit Gf wies der STEU teilweise keine Zusammenhänge (Austin, 2010), aber teilweise auch kleine ( $r = .10$ ) bis moderate ( $r = .35$ ) positive Zusammenhänge auf (Libbrecht & Lievens, 2012; MacCann, 2010).

Im Hinblick auf diskriminante Validitätsbelege, erwies sich der STEU als weitestgehend unabhängig von den Big-Five Persönlichkeitsfaktoren. Allein mit dem Faktor Offenheit für Erfahrungen ist der STEU gering (Libbrecht & Lievens, 2012) bis moderat

positiv (Ferguson & Austin, 2010; MacCann, 2010) korreliert. Zudem zeigt der STEU eine erwartungskonforme Nähe zu Messinstrumenten theoretisch ähnlicher Konstrukte, wie Affective Forecasting-Fähigkeit und sozial-kognitiver Theory of Mind (Hoerger et al., 2012; Ferguson & Austin, 2010). Insgesamt sprechen die bisherigen Belege für die Validität des STEU.

Mit dem STEU wurde die Domäne der EI erfolgreich um einen neuen Messansatz erweitert. Dabei besteht in den wenigen bisher verfügbaren Studien Heterogenität bezüglich der Reliabilität und der Zusammenhänge mit klassischer Intelligenz, die in zukünftigen Studien näher betrachtet werden sollte.

Der theoriebasierte Ansatz des STEU könnte dabei auch für die Entwicklung von Messverfahren in den übrigen Branches anregend sein. Aufgrund der Menge an teilweise widersprüchlichen Theorien zur Auslösung von Emotionen, ist das Problem des Scorings allerdings noch nicht final gelöst.

Mit dem STEU wird EU gemessen, indem den Testpersonen die Bewertungen von Situationen vorgegeben werden. Bei der Betrachtung des Tests ist auffallend, dass einige Items sehr knapp formuliert sind. Hierbei stellt sich die Frage, ob aus den Situationsbeschreibungen tatsächlich die Ausprägungen auf den Bewertungsdimensionen immer eindeutig feststellbar sind, was eine Voraussetzung für das Scoring anhand der Theorie ist. Zudem kann einschränkend angemerkt werden, dass im STEU die Vorhersage der Emotion zu einem Zeitpunkt getroffen wird, der unmittelbar vor dem Prozess der Auslösung der Emotion steht. Dabei ist es im Sinne der Appraisal-Theorien, dass es individuelle Unterschiede gibt, wie verschiedene Personen die gleiche Situation bewerten und dass die Fähigkeit diese Bewertungen vorherzusagen, ebenfalls einen wichtigen Aspekt von EU darstellen könnte, der sich von dem, was mit dem STEU gemessen wird, unterscheiden könnte.



### ***2.8.3.3 Test of Emotional Intelligence (TEMINT; Blickle, Momm, Liu, Witzki & Steinmayr, 2011)***

Der TEMINT ist ein deutschsprachiger SJT, in dem 12 Situationen aus Sicht realer Zielpersonen beschrieben werden (Amelang & Steinmayr, 2006). Die Testpersonen schätzen für jede Situation die Ausprägungen von zehn Emotionen ein, wobei die Testantworten anhand des Target-Scoring bewertet werden. Der TEMINT weist Schätzungen für Cronbachs  $\alpha$  auf, die bisher im Bereich von .72 bis .82 lagen (Amelang & Steinmayr, 2006; Blickle et al., 2009; Blickle et al., 2011). Der TEMINT weist erwartungskonform die höchste Beziehung zur Branch EU im MSCEIT auf, die allerdings nur von moderater Höhe ist (Blickle et al., 2011). Der TEMINT ist unabhängig von den Big-Five Persönlichkeitsfaktor, was als diskriminanter Beleg für seine Validität gewertet werden kann (Blickle et al., 2009). Nicht den Erwartungen entsprechend wies der TEMINT einen Zusammenhang nahe Null mit klassischer Intelligenz auf (Amelang & Steinmayr, 2006).

### ***2.8.3.4 Assessment of Children's Emotions Skill (ACES; Trentacosta & Izard, 2007)/ Emotional Knowledge Test - (EKT; Izard, Fine, Schultz, Mostow, Ackerman & Youngstrom, 2001)***

Beim ACES/EKT-Ansatz handelt es sich um eine Familie von Verfahren, in denen die emotionalen Fähigkeiten von Kindern im Kindergartenalter gemessen werden sollen. Der ACES ist das aktuellste Verfahren dieser Reihe. Der Untertest *Social Situations* des ACES umfasst 15 kurze Beschreibungen von Situationen, die eine Zielperson erlebt. Das Kind wird nach jeder Situation gefragt, welche Emotion aus einem Set von vorgegebenen Emotionen die Zielperson wahrscheinlich empfindet. Das Scoring der Items wurde hierbei von den Autorinnen und Autoren des Tests festgelegt.

### **2.8.3.5 Weitere Verfahren, die als Tests der EU vorgeschlagen wurden**

#### *Vocabulary of Emotions Test (VET; Costa, Faria & Takšić, 2011)*

Der VET ist ein Vocabulary Test mit Emotionsbegriffen, der bisher in kroatischer und spanischer Sprache vorliegt (Costa et al., 2011). Im VET werden den Testpersonen emotionsbezogene Zieladjektive vorgelegt, zusammen mit 6 weiteren Adjektiven, von denen jeweils dasjenige ausgewählt werden soll, welches mit der Bedeutung des Zielbegriffs am meisten übereinstimmt. Als Standard für das Scoring dient hier ein Wörterbuch. Studien, die die Validität des Tests belegen, liegen bisher noch nicht vor. Inhaltlich wird damit das Wissen über die Bedeutung von Emotionsbegriffen gemessen. Zwar wurde der JET als mögliches Maß für EU vorgeschlagen; ob dieser Test allerdings mehr als Wissen über Wortbedeutungen misst, ist fraglich.

#### *Judgment of Emotions Test (JET; Hoerger et al., 2012)*

Der JET wurde bisher als neues Forschungsinstrument in einer einzigen Studie erwähnt, in der dieser nicht im Fokus stand und eher am Rande behandelt wurde (Hoerger et al., 2012). Im JET werden 40 Items präsentiert, in denen die Testpersonen vorhersagen sollen, ob verschiedenen Variablen (z. B. Bildungsstand, Introversión) mit höher oder niedriger ausgeprägten Emotionen einhergehen (Hoerger et al., 2012). Als Grundlage für die Items und das Scoring dienten hierbei Forschungsergebnisse über Zusammenhänge dieser Variablen. Für den JET wurde ein Cronbachs  $\alpha$  von .81 ermittelt. Im Sinne konvergenter Belege, wies der JET eine Korrelation mit dem STEU von  $r = .59$  sowie eine positive Beziehung zu einem klassischen Wissenstest von  $r = .21$  auf. Bisher ist allerdings noch zu wenig über den JET bekannt (einschließlich des Item-Materials und der Scoring-Grundlage) als dass dieser als Test für EU bezeichnet werden kann.

*Levels of Emotional Awareness Scale (LEAS; Lane, Quinlan, Schwartz, Walker & Zeitlin, 1990)*

Im LEAS werden fiktive, interpersonale Situationen beschrieben, die von der Testperson erlebt werden könnte. Die Testperson soll einmal für sich selbst und einmal für eine andere Zielperson die emotionale Antwort in der gegebenen Situation im offenen Antwortformat beschreiben. Im LEAS werden die verbalen Beschreibungen der Testpersonen sprachlich anhand festgelegter Kriterien ausgewertet. Dabei wird bewertet, inwiefern die Testperson bei ihren Beschreibungen auf Emotionsbegriffe zurückgreift (hohes Ausmaß an Emotional Awareness) oder ob sie die emotionale Reaktion eher umschreibt und dabei keine Emotionen benennt (geringes Maß an Emotional Awareness).

Der LEAS wurde häufiger als mögliches Messinstrument im Rahmen von EI bzw. EU diskutiert (Mayer et al., 2008; Roberts et al., 2008). Zwar weist die Instruktion des LEAS, in Form der Vorhersage einer Emotion, Ähnlichkeit mit Tests für EU auf; allerdings wird im LEAS nicht die Akkuratheit der vorhergesagten Emotion bewertet, sondern das Ausmaß des Gebrauchs von Emotionsbegriffen. Aufgrund des Auswertungskriteriums, das die Art statt die Korrektheit der Beschreibung bewertet wird, erscheint der LEAS nicht zur Erfassung der EU geeignet zu sein.

#### **2.8.4 Selbstberichte der EI**

Der größte Teil der verfügbaren Messinstrumente, die zur Erfassung von EI vorgeschlagen wurden, sind Selbstberichtsverfahren. Selbstberichte der EI stellen fast ausschließlich Operationalisierungen von Modellen dar, die in den Bereich der Trait-EI oder Mixed Models fallen. Bereits in der Einleitung wurden einige Beispiele für Selbstberichtsverfahren genannt. Ein prominentes Beispiel für Selbstberichtsverfahren ist das EQ-i (Bar-On et al., 2000). Das EQ-i umfasst Items wie "I like helping people" oder "I know how to deal with upsetting problems" (S. 1116). Das erste Item-Beispiel veranschaulicht, dass mit dem EQ-i

Verhaltensdispositionen gemessen werden. Das zweite Item stellt eine Selbsteinschätzung einer emotionalen Fähigkeit dar.

Das EQ-i ist nahe Null mit klassischer Intelligenz korreliert (Farrelly & Austin, 2007; Livingstone & Day, 2005; O'Connor & Little, 2003), was für eine neue Form von Intelligenz nicht erwartet wird. Zudem weist das EQ-i sehr hohe Korrelationen mit etablierten Persönlichkeitskonzepten auf, wie mit Neurotizismus (-.72 für Frauen, -.62 für Männer) und Extraversion (.56 für Frauen; .52 für Männer) (Dawda & Hart, 2000). Ein ähnliches Muster an Korrelationen wird auch für andere Selbstberichte, wie die SEIS oder den TEIQue gefunden (siehe Roberts et al., 2008).

Diese Zusammenhänge lassen stark bezweifeln, dass mit Selbstberichten etwas anderes erfasst wird, als bereits etablierte Persönlichkeitskonstrukte. Wilhelm (2005) machte darauf aufmerksam, dass bei der Konstruktion einiger Selbstberichtsverfahren sogar auf Items etablierter Persönlichkeitstests zurückgegriffen wurde, so dass das damit erfasste Konstrukt sich kaum von etablierten Konstrukten unterscheiden kann.

Ebenso sind Selbsteinschätzungen von Fähigkeiten problematisch, da aus dem Bereich der klassischen Intelligenz bereits bekannt ist, dass diese nicht genügend mit tatsächlichen Fähigkeiten korreliert sind. In einer Meta-Analyse von Freund und Kasten (2012) über 41 Studien waren Leistungen in objektiven Intelligenztests mit Selbstberichten nur moderat positiv korreliert (mittlere Effektstärke von  $r = .33$ ). Wie zu erwarten, ist auch die tatsächliche Leistung im MSCEIT nur zu  $r = .20$  mit der selbsteingeschätzten EI korreliert (Sheldon et al., 2014). Zudem scheinen die Abweichungen zwischen tatsächlichen und selbstberichteten Fähigkeiten systematische Verzerrungen widerzuspiegeln, wobei sich Personen mit höheren Fähigkeiten tendenziell unterschätzen, während sich Personen mit geringen Fähigkeiten eher überschätzen (Kruger & Dunning, 1999). Dieser sogenannte Dunning-Kruger Effekt wurde auch im Bereich der EI gezeigt (Sheldon et al., 2014).

Selbstberichtsverfahren sind überdies anfällig gegenüber Faking. Day und Carroll (2008) zeigten in einer Studie, dass die Instruktion sich besser darzustellen, als man ist, zu einer Erhöhung der Testwerte im EQ-i aber nicht im MSCEIT führte. Das unter der Faking-Instruktion im Labor beobachtete Problem, scheint auch für tatsächliche Anwendungen von Selbstberichten der EI in High-Stakes Situationen eine Rolle zu spielen. Lievens et al. (2011) fanden höhere Testwerte für Bewerber auf eine Stelle im Vergleich zu einer Gruppe von Stelleninhabern.

Nicht verwunderlich ist die Tatsache, dass Leistungstests der EI und Selbstberichtsverfahren, wie der EQ-i nicht bis gering korreliert sind (siehe Brackett & Mayer, 2003). Selbstberichte scheinen viel mehr etwas zu messen, das der Persönlichkeitsdomäne zugeordnet werden kann. Für die Weiterentwicklung von Theorien innerhalb der Persönlichkeitstheorien könnten Selbstberichte der EI einen wichtigen Beitrag liefern (Roberts et al., 2008). Zur Messung von EI oder EU erscheinen Selbstberichte allerdings als ungeeignet.

## **2.9 Optimierung der Messung von EU mit dem Empathic Agent Paradigma**

Das EAP wurde von Ralf Schulze und Richard D. Roberts entwickelt (R. Schulze & R. D. Roberts, persönliche Kommunikation, 01. April, 2009) und ist insbesondere für die Erfassung von EU geeignet (Schulze et al., 2009). Testverfahren, die unter dem EAP konstruiert werden, sind durch ein Zwei-Phasen-Design charakterisiert. In der ersten Phase (der Aneignungsphase) soll die Testperson die emotionalen Reaktionen einer oder mehrerer Zielpersonen in einer bestimmten Situation beobachten. In der zweiten Phase (der Anwendungsphase) soll die Testperson die emotionalen Reaktionen der Zielperson in einer vergleichbaren Situation, auf Grundlage des erworbenen Wissens aus der Aneignungsphase, vorhersagen. Die Entwicklung des EAP wurde inspiriert durch das sogenannte Principal-Agent-Paradigma (PAP), das in der wirtschaftswissenschaftlichen Literatur verortet ist

(Huber, Ariely & Fischer, 2002). Im PAP wird die Fähigkeit einer Testperson (des Agent) gemessen, die Präferenzstruktur einer Zielperson (dem Principal) korrekt einschätzen zu können (Orchard et al., 2009). Der Agent beobachtet dabei zunächst ein Set von Entscheidungen des Principals zwischen Exemplaren mit Ausprägungen auf verschiedenen Attributen. Anschließend muss der Agent die Entscheidung des Principals in einer neuen Entscheidungssituation vorhersagen (Orchard et al., 2009). Hierzu muss der Principal auf Grundlage seiner Beobachtungen die Teilnutzenwerte der Attribute erschließen und sie auf die Entscheidung des Principals in der neuen Situation anwenden. Im Rahmen des EAP wurde dieses Zwei-Phasen-Design übernommen und auf emotionsbezogene Kontingenzen einer Zielperson bezogen. Diese emotionsbezogenen Kontingenzen können dabei Ketten von a) Ereignissen und Emotionen, b) Emotionen und Emotionen oder c) Emotionen und Verhaltensweisen darstellen (Schulze et al., 2009). Im Rahmen der Aneignungsphase besteht die Aufgabe der Testperson darin, diese emotionsbezogenen Kontingenzen der Zielpersonen zu erkennen und kurzfristig zu behalten. In der Anwendungsphase muss die Testperson verstehen, welche der zuvor erworbenen Kontingenzen auf die neue Situation anwendbar sind und muss diese korrekt abgerufen.

Die beiden Entwickler des EAP machten verschiedene Vorschläge darüber, wie Tests unter dem EAP gestaltet sein können (siehe z. B. Orchard et al., 2009). Für die vorliegende Arbeit sind zwei Aspekte dabei von größerer Bedeutung: zum einen machten die Entwickler darauf aufmerksam, dass der Grad, in dem die Fähigkeit zur Emotionswahrnehmung eine Rolle spielt, von Charakteristika der Aneignungsphase abhängen kann. In dieser Arbeit wurde versucht, den Einfluss von Emotionswahrnehmung unter Verwendung des EAP möglichst gering zu halten. Zum anderen erwähnten Schulze und Roberts, dass allgemeines Wissen über Emotionen, das bspw. konsistent mit Appraisal-Theorien ist, im EAP dann hilfreich ist, wenn es auf die Zielperson anwendbar ist (Orchard et al., 2009). In dieser Arbeit wurde

angenommen, dass der Grad, inwiefern sich eine Person typisch oder untypisch verhält, ein wichtiges, schwierigkeitsstiftendes Merkmal ist.

Mit dem EAP ist es möglich, Zielpersonen mit individuellen emotionsbezogenen Kontingenzen zu modellieren. Es muss dabei nicht auf theoretische Annahmen oder Expertenmeinungen darüber zurückgegriffen werden, wie die meisten Menschen in einer gegebenen Situation emotional reagieren würden. Es trägt hingegen dem Umstand Rechnung, dass es individuelle Unterschiede in emotionalen Reaktionen geben kann. Individuelle Unterschiede in den emotionsbezogenen Kontingenzen sind auch unter aktuellen Emotionstheorien, wie den bereits genannten Appraisal-Theorien, möglich. Hier wird explizit angenommen, dass verschiedene Personen auf die gleiche Situation unterschiedlich reagieren können, da sich die Bewertungen der Situation unterscheiden können (Roseman & Smith, 2001).

Das Zwei-Phasen Design des EAP stellt eine deutlich akkuratere Simulation des Prozesses des Verstehens von Emotionen dar, als andere Ansätze zur Messung von EU dies bieten (Orchard et al., 2009). Eine Vorhersage zukünftigen Verhaltens in der Anwendungsphase auf Grundlage vergangenen Verhaltens in der Aneignungsphase erscheint sinnvoll, wenn davon ausgegangen wird, dass Verhalten über die Zeit stabil ist. Das EAP erscheint dabei besonders mit Theorien der Persönlichkeit als Interaktion von Dispositionen und Situationen und daran anknüpfenden Theorien über die Personen-Wahrnehmung vereinbar (siehe Kammrath, Mendoza-Denton & Mischel, 2005). Hierbei wird angenommen, dass sich Dispositionen nicht nur in der Häufigkeit eines Verhaltens äußern, sondern sich auch in einem intraindividuellen Profil von Situationen-Verhaltens Beziehungen niederschlägt, die als *If-then-Profile* bezeichnet werden (Friesen & Kammrath, 2011). In Theorien über If-then-Profile wird angenommen, dass bspw. eine ängstliche Person nicht nur eine höhere Basis-Rate hat, mit Angst zu reagieren, sondern, dass es auch ein Profil an

Situationen gibt, die bei der Person Angst auslösen oder nicht (Friesen & Kammrath, 2011). In Bezug auf die Wahrnehmung von Personen haben Kammrath et al. (2005) gezeigt, dass sich Personen die If-then-Profile von Zielpersonen aneignen und diese zur Einschätzung der Zielpersonen verwenden. In diesem Zusammenhang wurden auch individuelle Unterschiede in der Fähigkeit diskutiert, (emotionsbezogene) If-then-Profile einer Zielperson erwerben zu können (Friesen & Kammrath, 2011), was hohe Ähnlichkeit mit EU besitzt.

Ein großer Vorteil des EAP liegt darin, dass mit ihm ein veridikales Scoring möglich wird. Die Korrektheit der Testantworten wird dabei im Prozess der Konstruktion der Aneignungsphase festgelegt. Das EAP erlaubt damit nicht nur ein systematisches Vorgehen bei der Konstruktion von Items, sondern auch ein begründetes Scoring-Rational. Gerade diese beiden Aspekte wurden von einigen Autoren als wichtigen Schritt zur Verbesserung der Messung von EI gefordert (siehe Schulze et al., 2007). Zudem ist es denkbar, mit dem EAP auch schwere Items zu erzeugen, was mit Emotionstheorien oder Gruppenurteilen als Scoring-Grundlage schwer erscheint. Eine potentielle Einschränkung des EAP könnte darin liegen, dass ein Zwei-Phasen Design, das damit Lernaufgaben ähnelt, auch zu einem gewissen Grad die Merkfähigkeit einer Person beanspruchen könnte. Dieser Aspekt muss bei der Konstruktion von EAP-Aufgaben und Untersuchungen zum EAP berücksichtigt werden muss.

Insgesamt hat das EAP aufgrund seiner hohen Flexibilität in der Ausgestaltung und der Möglichkeit eines veridikalen Scorings das Potential, die Messung im Bereich der EU bzw. EI deutlich zu verbessern.



## **3. Studie 1**

### **3.1 Hintergrund und Ziele**

Die erste Studie wurde mit dem Ziel durchgeführt, einen Test zur Erfassung von EU basierend auf dem EAP zu entwickeln, diesen hinsichtlich seiner psychometrischen Qualität zu untersuchen sowie erste Validitätsevidenz zu finden. Im ersten Schritt sollte ein konkretes Testdesign entwickelt werden, mit dem EU über das EAP operationalisiert werden kann, und anschließend sollten auf Grundlage dieses Designs Items konstruiert werden. Im zweiten Schritt bestand das Ziel darin, den entwickelten EAP-Test an einer Stichprobe zu erproben. Die mit der Stichprobe gewonnenen Daten sollten zum einen für Schätzungen der Reliabilität des Tests, der Selektion geeigneter Items und der Prüfung der angenommenen eindimensionalen Struktur herangezogen werden. Zum anderen diente die erste Stichprobe dazu, den EAP-Test im Rahmen seines nomologischen Netzwerks zu betrachten (siehe Cronbach & Meehl, 1955). In diesem Zusammenhang sollte geprüft werden, inwiefern aus der Theorie erwartete Beziehungen des EAP-Tests zu Operationalisierungen gleicher, ähnlicher und davon abzugrenzender Konstrukte gefunden werden können. Dazu wurden die Beziehungen des EAP-Tests zu einem anderen Test für EU, zu Maßen der klassischen Intelligenz und den Big-Five Persönlichkeitsfaktoren betrachtet.

Als neues Messinstrument für EU sollte der EAP-Test, im Sinne konvergenter Validitätsbelege, substantielle positive Beziehungen zu anderen Maßen von EU aufweisen. Aufgrund der mangelnden Verfügbarkeit von Messinstrumenten für EU, wurde im Rahmen der ersten Studie ein weiteres Messinstrument zur Erfassung von EU entwickelt: der Theory-Based Test of Emotional Understanding (TBEU). Der TBEU wurde theoriebasiert in Anlehnung an den STEU von MacCann und Roberts (2008) entwickelt. Damit stand ein Maß für EU zur Verfügung, dessen Konstruktion und Scoring auf einer anderen Methode als dem

EAP beruhte, so dass Zusammenhänge zwischen den Maßen von EU nicht allein auf identische Methoden zurückgeführt werden können (siehe Campbell & Fiske, 1959).

Zudem wurde erwartet, dass der EAP-Test positive Beziehungen zu Maßen der klassischen Intelligenz aufweist. Da es sich bei EU um eine Form von Intelligenz handeln soll, kann diese Erwartung aus der Beobachtung des Positive-Manifold-Phänomen abgeleitet werden (siehe z. B. Jensen, 1992). Die Wichtigkeit dieses konvergenten Validitätsbelegs für einen Test, der eine Form von Intelligenz erfassen soll, wurde in der Literatur mehrfach unterstrichen (z. B. Mayer et al., 1999). In Untersuchungen, die auf den bisherigen Tests zur Erfassung von EU basieren, wurden bisher in der Regel höhere Beziehungen zu Gc als zu Gf beobachtet (siehe Roberts et al., 2008). Diese Ergebnisse haben dazu geführt, dass EU von einigen als eine Form von Gc betrachtet wurde (z. B. Mayer et al., 2002). In Konzeptualisierungen wird EU allerdings auch als Fähigkeit zum Schlussfolgern über Emotionen beschrieben, was eine Nähe zu Gf impliziert (siehe Mayer et al., 2001). Ein Grund dafür, dass EU bislang eher Gc zugeordnet wurden, könnte in den wenigen verfügbaren Operationalisierungen von EU begründet liegen, die den Aspekt des Schlussfolgerns über Emotionen bislang vernachlässigen. Beispielsweise wird im Blends-Test des MEIS EU operationalisiert, indem Wissen über die Bedeutung von Emotionsbegriffen erfasst wird (siehe Mayer et al., 1999). Ebenso wird im Progressions-Test des MEIS das Wissen über allgemeingültige Regeln des Verlaufs von Emotionen erfragt (siehe Mayer et al., 1999). Die im EAP-Test gemessene Fähigkeit hingegen verlangt eher das Verstehen neuer Emotionskontingenzen, die noch kein Produkt vergangener Lernerfahrungen gewesen sein können. Für den EAP-Test wird daher eine höhere Beziehung zu Gf als zu Gc erwartet, was zudem nicht im Widerspruch mit der Konzeptualisierung von EU steht.

Gleichzeitig wurde erwartet, dass die mit dem EAP-Test gemessene Fähigkeit im Sinne diskriminanter Validitätsbelege von etablierten Intelligenzformen abgegrenzt werden

kann. Die Abgrenzung zu zwei Fähigkeiten erschien hier besonders entscheidend: zum einen von Merkfähigkeit, da der EAP-Test durch das Aneignungsphase-Anwendungsphase-Design das kurzfristige Behalten von Informationen erfordert; zum anderen von verbaler Intelligenz aufgrund der verbalen Präsentation der Items im EAP-Tests.

Zudem sollte der EAP-Test, als Maß für eine neue Form von Intelligenz, ein ähnliches Muster von Beziehungen zu den Big-Five Persönlichkeitsfaktoren aufweisen, wie es für bisherige Tests zur Erfassung von Intelligenz beobachtet wurde (siehe z. B. Ackerman & Heggestad, 1997). Demnach wird erwartet, dass der EAP-Test mit den Big-Five Persönlichkeitseigenschaften unkorreliert ist, mit Ausnahme des Faktors Offenheit für Erfahrungen, für den eine moderat positive Beziehung vorhergesagt werden kann.

Zusammenfassend wurde erwartet, dass der EAP-Test eine moderate bis hohe positive Beziehung zum TBEU aufweist, positive Beziehungen zu Maßen der klassischen Intelligenz zeigt, die wiederum nicht so hoch sind, dass der EAP-Test von klassischer Intelligenz nicht unterscheidbar ist, sowie mit den Big-Five Persönlichkeitsfaktoren unkorreliert ist, mit Ausnahme einer positiven, moderaten Beziehung zu Offenheit für Erfahrung. Zeigen sich die aus dem nomologischen Netzwerk erwarteten Beziehungen, können diese als konvergente und diskriminante Validitätsbelege für den EAP-Test bewertet werden.

## **3.2 Methode**

### **3.2.1 Entwicklung des EAP-Tests**

Im ersten Schritt der Entwicklung wurde das allgemeine Design des EAP-Tests festgelegt. Die Entwicklung des Designs wurde hauptsächlich durch drei Ziele geleitet: Das erste Ziel bestand darin, ein Testdesign zu erschaffen, das die Darstellung von Emotionskontingenzen einer Zielperson erlaubt, ohne dass die Testperson Emotionen in Gesichtsausdrücken, der Stimme oder der Körperhaltung identifizieren muss. Damit sollte verhindert werden, dass die

Erfassung von EU mit der Erfassung der Fähigkeit zur Emotionswahrnehmung konfundiert ist. Das zweite Ziel war, ein Design zu erstellen, mit dem der potentielle Einfluss von Merkfähigkeit reduziert wird. Hiermit sollte bspw. vermieden werden, dass eine Person trotz genügend hoher EU-Fähigkeit ein gegebenes Item zu lösen, dennoch eine falsche Antwort gibt, da sie die Namen der Zielpersonen falsch erinnert. Das dritte Ziel war, die Items in einer Art zu präsentieren, die überschaubar ist und für die Testperson einfach zu verstehen ist.

Im EAP-Test werden die Items im Rahmen von Gesprächen zwischen Personen präsentiert. Dies wurde als flexibler Ansatz betrachtet, da Gespräche zum einen die Möglichkeit bieten, Personen über Gründe von Emotionen kommunizieren zu lassen und zum anderen die Gespräche selbst Situationen darstellen können, die Emotionen auslösen können. Um das Design des EAP-Tests zu illustrieren, wird in Abbildung 3.1 der EAP-Test mit Screenshots aus der Aneignungsphase und Anwendungsphase illustriert. Die Aufgabe der Testperson ist, die Gespräche und emotionalen Reaktionen von zwei bis vier Zielpersonen zu beobachten. Die Zielpersonen werden mit Hilfe von Avataren repräsentiert, die an verschiedenen Ecken des Bildschirms platziert sind. Die Gespräche zwischen den Zielpersonen werden durch Sprechblasen symbolisiert. Die Emotionen, die eine Zielperson zu einem Zeitpunkt fühlt, werden als Begriffe unterhalb des Avatars angezeigt. Dabei kann eine Zielperson eine oder zwei Emotionen gleichzeitig fühlen, die wiederum drei Intensitätsstufen besitzen können (leicht, mittel und hoch). In Bezug auf das Ziel, die individuellen Unterschiede in der Fähigkeit der Emotionswahrnehmung zu kontrollieren, werden die Avatare ohne Gesichter und zudem in einer rudimentären und standardisierten Körperhaltung dargestellt. In der Aneignungsphase des EAP-Tests werden die Emotionskontingenzen auf zwei Arten präsentiert: Zum einen kann eine Zielperson direkt über eine Emotionskontingenz sprechen, so dass die relevante Information in der Sprechblase gegeben ist. Zum anderen kann eine emotionale Reaktion in einer spezifischen Situation aus dem Gesprächsverlauf resultieren. Die emotionale Reaktion wird dann unterhalb des Avatars angezeigt. In beiden

Fällen wird die emotionale Reaktion auf eindeutige Weise präsentiert, so dass die Fähigkeit zur Emotionserkennung nicht relevant ist, um die damit verbundenen Items zu lösen. Mit dem Ziel, die Unterscheidung und Wiedererkennung der Zielpersonen zu erleichtern, wurden die Avatare so erstellt, dass sie in Bezug auf die Kombination ihrer Haarfarbe, Frisur und Farbe der Kleidung einzigartig sind. Des Weiteren wurde darauf geachtet, dass die Namen der Zielpersonen innerhalb eines Szenarios gut zu unterscheiden sind. Die Namen werden außerdem dauerhaft oberhalb der Avatare angezeigt, so dass es insgesamt mehrere Möglichkeiten gibt, die Zielpersonen wiederzuerkennen.

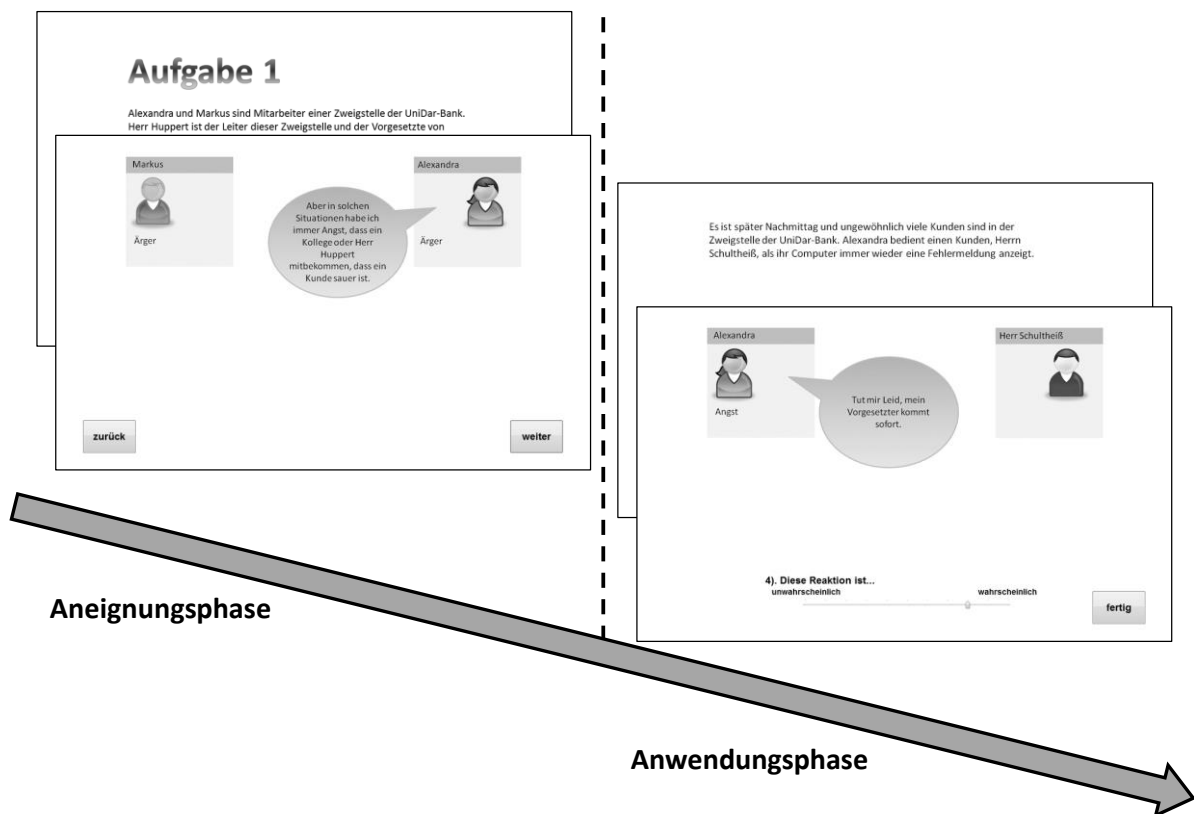


Abbildung 3.1. Illustration des EAP-Tests mit Screenshots aus der Aneignungs- und Anwendungsphase.

Der EAP-Test wurde mit der Programmier-Software Inquisit 3 umgesetzt. Die Gespräche werden als Sequenz von Einzelbildern, ähnlich einem Animationsfilm, präsentiert. Im Gegensatz zu einem animierten Film, kann die Testperson in selbst gewähltem Tempo durch den Test über die Vor- und Zurücktasten im unteren Bereich des Bildschirms navigieren. Das Tempo kann selbst gewählt werden, um den EAP-Test für Personen mit unterschiedlichen Lesegewohnheiten und -fähigkeiten möglichst fair zu gestalten. Innerhalb von Gesprächen darf die Testperson Bilder zurückzugehen, um Informationen bei Bedarf (wie beispielsweise bei einem versehentlichen Überspringen eines Bildes) erneut lesen zu können. Ist die Testperson in der Anwendungsphase, kann sie nicht mehr zurück in die Aneignungsphase wechseln. Damit soll verhindert werden, dass die Testperson zur Beantwortung eines Items die Aneignungsphase gezielt nach relevanten Informationen absucht.

Jede Aneignungsphase startet mit einer kurzen, textbasierten Beschreibung der Situation und der Zielpersonen. Die Aufgabe der Testperson ist, die Gespräche und emotionalen Reaktionen der Zielpersonen zu beobachten, wobei die Testperson zu diesem Zeitpunkt nicht weiß, welche Zielpersonen und welche ihrer emotionalen Kontingenzen für den Lösungsprozess relevant werden. Um die Testpersonen auf Veränderungen der emotionalen Zustände der Zielpersonen aufmerksam zu machen, werden Änderungen über einen Wechsel der Schriftfarbe signalisiert. Jedes Mal, wenn eine neue Emotion auftritt oder sich ein momentaner Zustand ändert, wird diese Änderung in roter Schriftfarbe eingeführt, während andauernde Zustände in schwarzer Schriftfarbe präsentiert werden. Fühlt eine Zielperson beispielsweise Ärger und später starken Ärger, dann wird der Begriff „stark“ in roter Schriftfarbe hinzugefügt und auf dem nächsten Bild in schwarzer Schriftfarbe dargestellt. Um Fehlinterpretationen der Schriftfarbe zu vermeiden, werden die Testpersonen zuvor in einer Instruktion über diese Funktion informiert.

In der Anwendungsphase des EAP-Tests werden einige der Zielpersonen in einer neuen Situation präsentiert. Die Anwendungsphase beginnt mit einer kurzen Situationsbeschreibung, der dann ein neues Gespräch folgt, das beobachtet werden soll. Die Situation in der Anwendungsphase ist so konstruiert, dass sie vergleichbar ist mit den Situationen, die in der Aneignungsphase als Auslöser für eine emotionale Reaktion definiert wurden. Zu einem bestimmten Zeitpunkt endet das Gespräch abrupt, so dass die emotionale Reaktion der Zielperson nicht mehr angezeigt wird. Die Aufgabe der Testperson ist es dann, die emotionale Reaktion vorherzusagen. Dazu werden nach einer kurzen Instruktion vier bis fünf mögliche emotionale Reaktionen der Zielperson nacheinander gezeigt. Für jede einzelne Reaktion soll die Testperson die Wahrscheinlichkeit einschätzen, unter Beachtung des erworbenen Wissens über die Emotionskontingenzen der Zielperson aus der Aneignungsphase. Die potentiellen Reaktionen variieren dabei hinsichtlich der Emotionen, die die Zielperson fühlt (das Emotionswort und die dargestellte Intensität unterhalb des Avatars) und ob die Emotion in der Antwort offen gezeigt oder verborgen wird (Verhalten in der Sprechblase). Die Testpersonen werden in der Instruktion darauf hingewiesen, dass eine Reaktion nur dann als wahrscheinlich einzuschätzen ist, wenn sowohl die Emotion als auch das Verhalten in der Sprechblase mit der Information über die Zielperson übereinstimmt. Die Testperson antwortet, indem sie einen Schieberegler auf einer 11-Punkte Skala von 0 „unwahrscheinlich“ bis 10 „wahrscheinlich“ bewegt. Die Default-Position ist der Mittelpunkt (5) der Skala. Um zu verhindern, dass die Testperson ungewollt ein Item überspringt, wird die Testperson angehalten den Schieberegler entweder auf die linke oder rechte Seite der Skala zu bewegen. Um dies technisch umsetzen zu können, wurde der Mittelpunkt als invalide Antwort definiert.

Zur Auswertung werden die Einschätzungen auf der Ratingskala zunächst am Mittelpunkt in die Antworten „unwahrscheinlich“ (Antworten von 0 bis 4) oder

„wahrscheinlich“ (Antworten 6 bis 10) dichotomisiert. Danach werden die dichotomisierten Einschätzungen als entweder inkorrekt (0 Punkte) oder korrekt (1 Punkt) bewertet.

Obwohl die Personen ihre Einschätzungen auf 10 möglichen Stufen vornehmen, werden diese nur in dichotomisierter Form bewertet. Es wird davon ausgegangen, dass in den erstellten Items im EAP-Test keine feinere Differenzierung von Wahrscheinlichkeiten möglich ist. Emotionale Reaktionen mit einer Wahrscheinlichkeit anders als 0 oder 1 hätten nur dann eindeutig spezifiziert werden können, wenn die Ziele in mehr als einer Situation gezeigt worden wären, was die Aneignungsphasen deutlich verlängert hätte. Nichtsdestotrotz wird den Testpersonen gestattet, ihre subjektiv empfundenen Wahrscheinlichkeiten in ihrer Einschätzung mit auszudrücken, die nicht unbedingt 0 oder 1 sein müssen. Das bedeutet bspw., dass eine Testperson ihre Einschätzung zum Ausdruck bringen kann, dass eine emotionale Reaktion unwahrscheinlicher als eine andere ist, selbst wenn die Aneignungsphase diese Unterscheidung nicht erlaubt.

Im zweiten Schritt der Konstruktion wurden 6 Szenarien für den EAP-Test erstellt. Innerhalb dieser Szenarien wurden insgesamt 15 Itemgruppen konstruiert, die je 5 bis 6 Items enthalten, die sich jeweils auf die gleiche Emotionskontingenz beziehen. Bei der Generierung der Items wurden alle Emotionskontingenzen spezifiziert und dokumentiert, die beim Lösungsprozess der Items korrekt angewendet werden müssen. Damit sollte sichergestellt werden, dass das Scoring im EAP-Test tatsächlich auf Grundlage des EAP begründet werden kann. Als Grundlage für die Wahl der Emotionen, die Inhalte der Items werden sollten, wurde die von Shaver, Schwartz, Kirson und O'Connor (1987) vorgeschlagene Taxonomie von Emotionsbegriffen herangezogen, um sicherzustellen, dass es sich um akzeptierte Emotionsbegriffe handelt.

Ein besonderes Ziel bei der Konstruktion der Items bestand darin, die Schwierigkeit der Items zu variieren und insbesondere auch schwierige Items zu erzeugen. Zeidner,



Matthews und Roberts (2001) haben darauf hingewiesen, dass die Identifikation schwieriger Items problematisch sein kann, wenn die mehrheitlich gewählte Antwort als richtige Antwort definiert wird, wie es im Consensus Scoring der Fall ist. Auch bei theoriebasierten Ansätzen, wie dem STEU (MacCann & Roberts, 2008) ist ein ähnliches Problem naheliegend, da die dahinterliegende Theorie teilweise auf Befragung und somit ebenfalls auf Mehrheitsurteilen beruht (siehe Roseman, 2001). Bei der Konstruktion des EAP-Tests wurde versucht, über mehrere Merkmale der Aufgaben die Schwierigkeit zu erhöhen. Vermutet wurde, dass eine hohe Anzahl von Emotionskontingenzen sowie ungewöhnliche Emotionskontingenzen die Schwierigkeit von Items erhöhen.

Die Durchführung des EAP-Tests startet mit einer umfassenden Instruktion, die ein komplettes Itembeispiel beinhaltet. Damit soll sichergestellt werden, dass die Aufgabe im EAP-Test von allen Getesteten verstanden wird. Um die Instruktion standardisiert durchzuführen, werden die Instruktionstexte zusätzlich über Kopfhörer dargeboten. Zudem werden einige Informationen auf diesem Weg ergänzt, um den Umfang an zu lesenden Instruktionstexten für die Testpersonen zu reduzieren. In der Instruktion wird betont, dass eine Antwort nur dann korrekt ist, wenn sie mit dem erworbenen Wissen über die Zielperson aus der Aneignungsphase übereinstimmt.

### **3.2.2 Entwicklung des Theory Based Test of Emotional Understanding (TBEU)**

Der TBEU wurde in dieser Studie als zweiter Test zur Erfassung von EU entwickelt, um ein weiteres, mit einer anderen Methode konstruiertes, Verfahren zur Verfügung stehen zu haben. Die Konstruktion und das Scoring des TBEU wurde in Anlehnung an den STEU von MacCann und Roberts (2008) auf Grundlage von Rosemans Appraisal Theorie vollzogen. Zwar basieren TBEU und STEU auf der gleichen Emotionstheorie, dennoch unterscheiden sich die beiden Tests in wesentlichen Punkten voneinander. Im TBEU werden die Situationen direkt aus der Perspektive der Zielpersonen in Form von direkter Rede beschrieben. Damit

sollte sichergestellt werden, dass die kognitiven Bewertungen der Zielpersonen, auf deren Grundlage die Testpersonen über die Emotionen schließen sollen, eindeutig spezifiziert werden können.

Um die Aufgabe für die Testpersonen abwechslungsreicher zu gestalten, wurden zwei verschiedene Aufgabentypen erstellt, die sich im Design, aber nicht in der dahinterliegenden Anforderung unterscheiden sollten. In einem Teil der Aufgaben enthält der Itemstamm den Bericht einer Zielperson über ein Ereignis. Die Aufgabe der Testperson besteht darin, für eine Reihe von Emotionen nacheinander anzugeben, ob diese wahrscheinlich oder unwahrscheinlich durch diese Bewertung der Situation in der Zielperson ausgelöst werden. In einem anderen Teil der Items wird im Itemstamm eine Emotion oder ein Verlauf von Emotionen der Zielperson vorgegeben. Dazu wird der Testperson eine Reihe von Ereignissen nacheinander vorgelegt, für die sie jeweils einschätzen soll, ob sie wahrscheinliche oder unwahrscheinliche Ursachen für die Emotionen in der Zielperson sind.

Für einen ersten Itempool wurden 25 Aufgabenbündel erstellt, in die jeweils 5 bis 6 Items genestet sind. Insgesamt umfasst der TBEU 138 Items. Innerhalb jedes Aufgabenbündels können mehrere emotionale Reaktionen bzw. beschriebene Ereignisse wahrscheinlich sein. Hierzu wurden bei einigen Aufgaben die Ereignisse so beschrieben, dass nicht alle Bewertungsdimensionen festgelegt wurden. Bei diesen Aufgaben muss die Testperson anhand der unvollständigen Informationen über das Bewertungsmuster eine Einschätzung vornehmen. Die Antworten auf den 138 Items werden jeweils als korrekt mit einem Punkt bewertet, wenn sie mit der aus der Theorie abgeleiteten Vorhersage übereinstimmen und andernfalls als falsch mit 0 Punkten bewertet.

### **3.2.3 Umgang mit Testlet-Effekten**

Lokale stochastische Unabhängigkeit der Items ist eine Voraussetzung, die in verschiedenen Testmodellen getroffen werden muss (Lord & Novick, 1974). Lokale stochastische

Unabhängigkeit der Items liegt dann vor, wenn Zusammenhänge zwischen Items allein auf die latenten Variablen zurückgeführt werden können (Lord & Novick, 1974). In vielen Fällen muss bezweifelt werden, dass die in der Item-Response-Theory verlangte Annahme lokaler stochastischer Unabhängigkeit oder die in der klassischen Testtheorie verlangte Unkorreliertheit der Fehler eine realistische Annahme darstellt (Lee, Dunbar & Frisbie, 2001). Ein Beispiel hierfür sind Tests, bei denen sich Subgruppen von Items auf den gleichen Gegenstand, wie eine Textpassage, die zur Beantwortung der Fragen relevante Informationen liefert, beziehen (Wainer, Bradlow & Wang, 2007). Sowohl der EAP-Test als auch der TBEU sind so gestaltet, dass mehrere Einschätzungen mit der gleichen Situationsbeschreibung, dem gleichen Szenario oder der gleichen Emotionskontingenz gepaart sind. Mit dem Problem der potentiellen Gefahr der Verletzung lokaler Unabhängigkeit, wird auf unterschiedliche Weise umgegangen.

In vielen Fällen wird das Problem schlichtweg ignoriert, was z. B. von Wainer et al. (2007) kritisiert wurde. Beispielsweise gibt es Hinweise darauf, dass die Reliabilität überschätzt wird, wenn abhängige Items als unabhängige Items in Testmodellen behandelt werden (z. B. Lee et al., 2001; Sireci, Thissen & Wainer, 1991; Wang, Bradlow & Wainer, 2002).

Eine andere Möglichkeit besteht darin, das Problem der Abhängigkeit durch das Testdesign zu vermeiden, indem immer nur je ein Item auf eine Textpassage bezogen wird (Wainer et al., 2007). Effizienter ist es allerdings, wenn mehrere Items auf eine Textpassage bezogen werden, da so mehr Information erhoben werden kann (Wainer et al., 2007). Vor allem für Tests, bei denen das Lesen relativ umfangreicher Situationsbeschreibungen erforderlich ist, wie es beim EAP-Test und TBEU der Fall ist, stellt der Vorteil der Effizienz ein entscheidendes Kriterium für ein Testdesign mit mehreren Abfragen zum gleichen Gegenstand dar.

Eine weitere Möglichkeit im Umgang mit dieser Problematik besteht darin, die Abhängigkeit der Items im Rahmen von Testmodellen zu berücksichtigen (Wainer et al., 2007). In diesem Zusammenhang wurde das Konzept der Testlets vorgeschlagen. Das Testlet-Konzept wurde ursprünglich von Wainer und Kiely (1987) im Kontext des Computergestützten Adaptiven Testens eingeführt und später auf sämtliche Kontexte, bei denen die Abhängigkeit von Items ein Problem darstellt, verallgemeinert. Lee, Brennan und Frisbie (2000, S. 10) definierten Testlets als „a subset of the items in a test form that is treated as a measurement unit in test construction, administration, and/or scoring“. Testlets können sowohl aus mehreren abhängigen Items bestehen als auch einzelne Items repräsentieren, die von allen anderen unabhängig sind (Lee et al., 2000).

Mit Testlets kann auf unterschiedliche Weise umgegangen werden. Eine Möglichkeit besteht darin, die Items innerhalb eines Testlets, z. B. durch Summenbildung, zu kombinieren und anstelle einzelner Items als Untersuchungseinheit zu verwenden (z. B. Lee et al., 2000). Reliabilitätsschätzungen können dann über die Testlets anstatt der Items berichtet werden, wodurch konservativere Schätzungen der Reliabilität möglich sind (Wang et al., 2002). Des Weiteren wurden spezielle Testmodelle vorgeschlagen, in denen die Abhängigkeit von Items modelliert und geschätzt werden kann (Wainer et al., 2007). Hierzu wurden im Rahmen der Item-Response Theory eine Reihe sogenannter Testlet-Modelle vorgeschlagen.

Wang et al. (2002) haben ein Testlet-Modell entwickelt, das sowohl für dichotome als auch polytome Items geeignet ist und auf dem 3-parametrischen logistischen Modell beruht. In diesem Modell bezeichnet  $\gamma_{id(j)}$  einen zusätzlichen Parameter für den Testlet-Effekt von Item  $j$ , das in Testlet  $d(j)$  genestet ist, auf Person  $i$  (Wang et al., 2002). Die Varianz des Testlet-Effekts,  $\sigma_{d(j)}^2$ , die das Ausmaß der Abhängigkeit der Items innerhalb eines Testlets darstellt, ist für die Testlets spezifisch und kann somit zwischen den Testlets variieren (Wang et al., 2002).

Sowohl für den EAP-Test als auch für den TBEU wurde vermutet, dass Testlet-Effekte beobachtet werden können. In dieser Studie wurden diese potentiellen Testlet-Effekte auf zwei Arten berücksichtigt. Zum einen wurden Subgruppen von Items, die sich auf den gleichen Gegenstand beziehen, zu Testlets zusammengefügt. Die Testlets werden dann als Analyseeinheit für Reliabilitätsschätzungen und Strukturgleichungsmodelle verwendet.

Zum anderen wird das Ausmaß der Abhängigkeit der Items untersucht, indem das Modell von Wang et al. (2002) auf die Daten angepasst wird und die Varianzen der Testlet-Effekte geschätzt werden. Hierzu wurde das R-Paket Supplementary Item Response Theory Models (sirt; Robitzsch, 2014) verwendet.

### **3.2.4 Stichprobe**

Die Stichprobe umfasste 210 Personen (150 Frauen, 8 Personen haben ihr Geschlecht nicht angegeben). Das Alter der Teilnehmerinnen und Teilnehmer reichte von 16 bis 59 Jahren, mit einem Durchschnittsalter von 23.79 ( $SD = 8.56$ ) Jahren. Der größte Teil der Stichprobe ( $n = 82$ ) wurde an der Bergischen Universität Wuppertal angeworben und setzte sich hauptsächlich aus Studierenden zusammen. Des Weiteren wurden Schülerinnen und Schüler aus insgesamt vier Klassen zweier Gymnasien getestet ( $n = 73$ ). Darüber hinaus wurde die Stichprobe durch Teilnehmerinnen und Teilnehmer ergänzt, die im privaten und/oder nebenberuflichen Kontext von zwei Forschungspraktikantinnen angeworben wurden ( $n = 55$ ). 88.6% der Stichprobe gaben an, Deutsch als Muttersprache zu haben. Alle Teilnehmerinnen und Teilnehmer, die nicht Deutsch als Muttersprache hatte, sprachen seit mindestens 10 Jahren Deutsch.

### **3.2.5 Messinstrumente**

#### ***EU***

*EAP-Test.* Der im Rahmen der ersten Studie entwickelte EAP-Test wurde von den Teilnehmerinnen und Teilnehmern bearbeitet.

*TBEU*. Der ebenfalls im Rahmen der ersten Studie entwickelte *TBEU* wurde den Testpersonen vorgelegt.

### ***Klassische Intelligenz***

Klassische Intelligenz wurde mit vier Aufgabengruppen aus dem Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R; Liepmann, Beauducel, Brocke & Amthauer, 2006) gemessen. Die Theorie, die dem I-S-T 2000 R zugrunde liegt, stellt eine Konvergenz verschiedener Intelligenztheorien dar und beinhaltet unter anderem die Unterscheidung von Gf und Gc auf Ebene der Generalfaktoren. Die Aufgabengruppen zur Erfassung von Gf und Gc werden jeweils über verbale, figurale und numerische Inhalte operationalisiert. Für den I-S-T 2000 R liegen, abhängig von der Skala, zwei oder drei Parallelförmigkeiten vor. Für die hier verwendeten Aufgabengruppen wurde jeweils die Parallelförmigkeit ausgewählt, für die im Manual die höchste Reliabilitätsschätzung berichtet wurde. Alle Untertests wurden mit der, im Manual angegebenen zeitlichen Begrenzung durchgeführt.

*Merkfähigkeit verbal (Form A)*. Bei den Merkaufgaben verbal (Me) werden den Testpersonen 13 Begriffe vorgelegt, die insgesamt fünf Oberkategorien zugeordnet sind. Nach der Einprägungsphase werden der Testperson die Anfangsbuchstaben von zehn dieser Begriffe vorgelegt. Die Testperson muss anhand des Anfangsbuchstabens den Begriff abrufen und dann für diesen Begriff die zugeordnete Oberkategorie aus fünf vorgegebenen ankreuzen. Die Aufgabengruppe Merkaufgaben verbal umfasst insgesamt 10 Items.

*Analogien (Form C)*. Die Aufgabengruppe Analogien (An) besteht aus Items, in denen der Testperson jeweils drei Begriffe vorgelegt werden. Zwischen den beiden ersten Begriffen besteht eine Beziehung, die von der Testperson erkannt werden muss und auf den dritten Begriff angewendet werden muss. Dazu wählt die Testperson einen vierten Begriff aus fünf vorgegebenen Alternativen, der in einer ähnlichen Beziehung zum dritten Begriff steht, wie der erste zum zweiten Begriff. Insgesamt umfasst diese Aufgabengruppe 20 Items.

*Matrizen (Form C).* In der Aufgabengruppe Matrizen (Ma) werden Figuren vorgegeben, die einer bestimmten Regel folgenden aufgebaut sind. Die Aufgabe der Testperson ist, eine fehlende Figur regelkonform zu ergänzen, indem sie die entsprechende Figur aus fünf vorgegebenen Alternativen auswählt. Diese Aufgabengruppe besteht aus 20 Items. Da Matrizen-Tests eine sehr zentrale Position innerhalb von Verfahren zur Messung von Gf einnehmen (siehe Carpenter, Just & Shell, 1990), wurde dieser als Marker für Gf betrachtet.

*Wissen verbal (Form A).* Im Wissenstest (Wi) werden Multiple-Choice Fragen zu verschiedenen Wissensgebieten gestellt. Diese umfassen Geografie/Geschichte, Wirtschaft, Kunst/Kultur, Mathematik, Naturwissenschaften und Alltag. Zu jeder Frage werden der Testperson fünf Antworten vorgelegt, aus der die Testperson die richtige Alternative auswählen soll. Der Wissenstest stellte in dieser Studie einen Marker für Gc dar. Da die Emotionskontingenzen im EAP-Test verbal kodiert sind, erschienen die verbal kodierten Items des Wissenstests in dieser Studie von besonderem Interesse. Für diese Untersuchung wurden daher die 28 verbal kodierten Items des Wissenstest ausgewählt.

### ***Big-Five Persönlichkeitseigenschaften***

*Openness Conscientiousness Extraversion Agreeableness Neuroticism Index Condensed.* Die Big-Five Persönlichkeitseigenschaften wurden mit der deutschen Fassung des Openness Conscientiousness Extraversion Agreeableness Neuroticism Index Condensed (OCEANIC; Schulze & Roberts, 2006) erfasst. Als Ausgangspunkt für die Konstruktion des OCEANIC diente das Modell von Christal sowie das Trait Self Description Inventory (TSDI) (Schulze & Roberts, 2006). Der OCEANIC besteht aus Items, in deren Kern entweder Verhaltensaussagen (wie z. B. „Wenn ich etwas anfangen, dann arbeite ich, bis es zu meiner Zufriedenheit erledigt ist.“) oder Eigenschaftsbegriffe (wie z. B. „Ich bin ein Perfektionist.“) stehen. Während im TSDI Verhaltensaussagen und Eigenschaftsbegriffe über unterschiedliche Antwortformate abgefragt wurden, werden diese im OCEANIC über ein einheitliches

Antwortformat mit leicht adaptierten Aussagen erfragt. In der deutschsprachigen Fassung wird eine 6-stufige Likert-Skala verwendet, bei der „Ablehnung“ und „Zustimmung“ mit den jeweils drei Abstufungen „stark“, „mittel“ und „leicht“ differenziert werden. Die vorliegende deutsche Fassung des OCEANIC ist mit 45 Items sehr ökonomisch. Im Vergleich zu anderen Kurzskalen hebt den OCEANIC positiv hervor, dass bei seiner Konstruktion die qualitativ besten Items gewählt wurden und diese gleichzeitig ein breites Spektrum von Facetten systematisch abdecken (Schulze & Roberts, 2006). Im Folgenden werden die Big-Five entsprechend ihrer englischen Begriffe abgekürzt: Offenheit für Erfahrung (O), Gewissenhaftigkeit (C), Extraversion (E), Verträglichkeit (A) und Neurotizismus (N).

### 3.2.6 Durchführung

Die an der Bergischen Universität Wuppertal angeworbenen Testpersonen wurden in einem Laborraum in Gruppengrößen bis maximal acht Testpersonen von ein bis zwei Testleiterinnen und Testleitern getestet. Die Schülerinnen und Schüler wurden in ihren Klassenräumen in Gruppen bis 21 Personen von drei Testleiterinnen getestet. Im privaten oder nebenberuflichen Kontext angeworbene Teilnehmerinnen und Teilnehmer wurden entweder zu Hause oder in einem ruhigen Raum am Arbeitsplatz in Gruppen bis zu sechs Personen von ein bis zwei Testleiterinnen getestet. Bei allen Testungen wurde auf eine störungsfreien Umgebung und ein standardisiertes Vorgehen geachtet. In Tabelle 3.1 wird ein Überblick über den Ablauf der Testung gezeigt.

Tabelle 3.1

#### *Übersicht der in Studie 1 verwendeten Durchführungsdesigns*

Design	Dauer in Stunden	Darreichungsform	Abfolge der Tests
Vollständiges Design	ca. 2 – 2.5	Testheft	1. I-S-T 2000 R: Me 2. I-S-T 2000 R: An



Design	Dauer in Stunden	Darreichungsform	Abfolge der Tests
			3. I-S-T 2000 R: Ma
			4. I-S-T 2000 R: Wi (5 Minuten Pause)
		Computer	5. EAP-Test
		Testheft	6. TBEU
			7. Demo
			8. OCEANIC
Kurzes Design 1	ca. 1 – 1.5	Testheft	1. I-S-T 2000 R: Me
			2. I-S-T 2000 R: Ma
			3. I-S-T 2000 R: Wi
			4. Demo
		Computer	5. EAP-Test
Kurzes Design 2	ca. 1 – 1.5	Computer	1. EAP-Test
		Testheft	2. TBEU
			3. Demo
			4. OCEANIC

*Anmerkungen.* Demo = Demografische Fragen.

Aufgrund von Begrenzungen der zur Verfügung gestellten Testungszeiten an Schulen und am Arbeitsplatz, konnte nicht immer das vollständige Untersuchungsdesign durchgeführt werden. In diesen Fällen kamen zwei verkürzte Designs zum Einsatz, wobei der EAP-Test entweder mit einem Teil der Maße für klassische Intelligenz oder gemeinsam mit TBEU und OCEANIC dargeboten wurde. Die Dauer der Testungen konnte zwischen den Testpersonen variieren, da der EAP-Test, TBEU und OCEANIC im selbstgewählten Tempo bearbeitet wurden. Wie aus der Tabelle zu erkennen, wurde der EAP-Test an Computern präsentiert und

alle anderen Verfahren in Papier-und-Bleistift-Form in Testheften vorgelegt. Zur Durchführung des EAP-Tests wurden den Testpersonen Notebooks mit einer Bildschirmdiagonale von 15.6 Zoll vorgelegt sowie eine Maus und Kopfhörer bereitgelegt.

### 3.3 Ergebnisse

#### 3.3.1 Itemselektion des EAP-Tests

Zunächst wurde eine Itemselektion des EAP-Tests anhand von Trennschärfe- und Schwierigkeitsanalysen vorgenommen. Die Entscheidung über die Qualität der Items erfolgte auf Item- und nicht auf Testlet-Ebene, damit Items mit schlechten psychometrischen Eigenschaften innerhalb eines Testlets identifiziert werden konnten, ohne gleichzeitig qualitativ gute Items zu verwerfen.

Die Itemselektion erfolgte in mehreren Schritten. Im ersten Schritt wurden alle Items mit negativer Trennschärfe ausgeschlossen, was auf fünf Items zutraf, sowie Items mit Leichtigkeiten von  $\leq .05$  oder  $\geq .95$ , was für ein Item zutreffend war. Im zweiten Schritt wurden die Trennschärfen der Items über das reduzierte Set geschätzt und Items mit Trennschärfen  $\leq .05$  entfernt, wodurch sechs weitere Items ausgeschlossen wurden. Im dritten Schritt wurde das Vorgehen vom zweiten Schritt wiederholt und ein weiteres Item wurde ausgeschlossen. Im Anhang A, Tabelle A.1 werden die Trennschärfen und Leichtigkeiten der Items präsentiert. Das finale Set umfasste 63 Items, deren Trennschärfen von .05 bis .46 ( $M = .20$ ) reichten und deren Leichtigkeiten von .09 bis .94 ( $M = .61$ ) reichten. Wie in Tabelle A.1 entnommen werden kann, sind eine Reihe sehr leichter Items enthalten. Obwohl diese relativ wenig Information liefern können, wurde keine weitere Selektion nach Leichtigkeit vorgenommen, da dies die Reliabilität des EAP-Tests eher verschlechterte. Auf Itemebene lagen die Reliabilitätsschätzung für den ursprünglichen Itempool bei  $\alpha = .73$  und für die finale Itemauswahl bei  $\alpha = .77$ .

Anschließend wurden die ausgewählten Items zu Testlets zusammengefügt. Hierzu wurden zum einen aus den selektierten Items 14 Itembündel gebildet, die sich jeweils auf die gleiche Emotionskontingenz beziehen. Zum anderen wurden sechs Itembündel gebildet, in denen jeweils alle Items eines Szenarios zusammengefasst sind.

### **3.3.2 Itemselektion des TBEU**

Für den TBEU wurde ebenfalls schrittweise eine Itemselektion anhand von Trennschärfe- und Schwierigkeitsanalysen durchgeführt. Im ersten Schritt wurden alle Items mit negativen Trennschärfen ausgeschlossen, was auf 39 Items zutraf. Außerdem wurden in diesem Schritt zwei Items ohne Varianz entfernt, von denen eines eine Leichtigkeit von 1 und eines eine Leichtigkeit von 0 aufwies. Ein weiteres Item mit einer Leichtigkeit von  $> .99$  wurde außerdem ausgeschlossen. Da im ersten Schritt bereits eine große Anzahl von Items ausgeschlossen wurde, wurde keine weitere Selektion anhand von Schwierigkeiten vorgenommen. Im zweiten Schritt wurden die Trennschärfen für die verbliebene Auswahl von Items geschätzt und Items mit Trennschärfen von  $\leq .05$  ausgeschlossen, wodurch 11 Items entfernt wurden. Das Vorgehen aus dem zweiten Schritt wurde im dritten Schritt wiederholt und drei weitere Items wurden ausgeschlossen. Nach dem dritten Schritt lagen keine weiteren Items mit negativen Trennschärfen vor. Im Anhang B können der Tabelle B.1 die Trennschärfen und Leichtigkeiten der Items entnommen werden. Die Trennschärfen der Items reichten von .06 bis .45 ( $M = .22$ ) und die Leichtigkeiten der Items lagen zwischen .03 und .99 ( $M = .69$ ). Die Schätzung von Cronbachs  $\alpha$  auf Ebene der Items verbesserte sich durch die Itemselektion von .58 auf .84. Anschließend wurden die verbliebenen 82 Items zu 25 Testlets entsprechend der erstellten Itembündel zusammengefügt, die den weiteren Analysen in Studie 1 als Analyseeinheit zugrunde lagen.

### 3.3.3 Testlet-Effekte

Im EAP-Test wurden zwei verschiedene mögliche Quellen für Abhängigkeiten der Items indentifiziert. Zum einen erschien es naheliegend, dass die Items, die mit einer zu verstehenden Emotionskontingenz gepaart sind, nicht unabhängig voneinander sind. Zum anderen könnten auch Itemgruppen innerhalb der Szenarien aufgrund der gemeinsamen Aneignungsphase abhängig voneinander sein. Daher wurden zwei verschiedene Schätzungen für die Testlet-Effekte vorgenommen, die diese möglichen Quellen berücksichtigen.

Tabelle 3.2 zeigt die Punktschätzungen und Konfidenzintervalle für den Testlet-Effekt  $\text{var}(\gamma)$  innerhalb der 14 Itemgruppen zu je einer Emotionskontingenz, in die insgesamt 63 dichotome Items genestet sind. Wie aus der Tabelle zu erkennen ist, sind sämtliche Testlet-Effekte innerhalb der Itemgruppen größer Null. Zudem ist zu erkennen, dass die Schätzungen für die Testlet-Effekte zwischen den Itemgruppen deutlich variieren.

Tabelle 3.2

*Schätzungen der Testlet-Effekte im EAP-Test: 63 Items genestet in 14 Itemgruppen*

EAP Itemgruppe	Anzahl Items	Testlet Effekt $\text{var}(\gamma)$ [90%-Konfidenzintervall]
EAP1_1	4	0.35 [0.14 – 0.64]
EAP2_1	4	0.48 [0.25 – 0.78]
EAP2_2	4	1.72 [0.88 – 2.89]
EAP2_3	3	0.38 [0.23 – 0.56]
EAP3_1	5	2.62 [2.14 – 3.02]
EAP3_2	6	1.54 [1.14 – 1.84]
EAP3_3	6	2.72 [1.64 – 3.65]
EAP4_1	5	0.69 [0.28 – 1.15]

EAP Itemgruppe	Anzahl Items	Testlet Effekt $\text{var}(\gamma)$ [90%-Konfidenzintervall]
EAP4_2	4	0.30 [0.13 – 0.56]
EAP5_1	5	0.82 [0.54 – 1.06]
EAP5_2	6	0.79 [0.39 – 1.25]
EAP6_1	5	2.16 [1.42 – 2.72]
EAP6_2	3	0.37 [0.20 – 0.65]
EAP6_3	3	0.43 [0.19 – 0.79]

*Anmerkungen.* Schätzungen der Testlet-Effekte im Rahmen des 3PNO Testlet-Modells nach Wang et al. (2002). EAP1\_1 bis EAP6\_3 bezeichnen 14 Gruppen von je 3 bis 6 Items, die sich auf die gleiche Emotionskontingenz beziehen.

In Tabelle 3.3 werden die Schätzungen für die Testlet-Effekte von 14 Itemgruppen innerhalb der 6 Szenarien präsentiert. Wie aus der Tabelle ersichtlich wird, fallen die Schätzungen für den Testlet-Effekt geringer aus. Allerdings sind auch hier sämtliche Schätzungen für die Testlet-Effekte größer als Null. Da auch innerhalb der Szenarien Testlet-Effekte zwischen den Itemgruppen bestehen, wurden für alle weiteren Analysen mit dem EAP-Test die Szenarien als Analyseeinheit verwendet.

Tabelle 3.3

*Schätzungen der Testlet-Effekte im EAP-Test: 13 Itemgruppen genestet in 5 Szenarien*

Szenario	Anzahl Itemgruppen	Testlet Effekt $\text{var}(\gamma)$ [90%-Konfidenzintervall]
EAP1	1	-

Szenario	Anzahl Itemgruppen	Testlet Effekt $\vargamma$ [90%-Konfidenzintervall]
EAP2	3	0.48 [0.22 – 0.85]
EAP3	3	0.37 [0.19 – 0.78]
EAP4	2	0.36 [0.23 – 0.48]
EAP5	2	0.41 [0.17 – 0.84]
EAP6	3	0.31 [0.15 – 0.54]

*Anmerkungen.* Schätzungen für Testlet-Effekte im Rahmen des 3PNO Testlet-Modells nach Wang et al. (2002). EAP1 bis EAP6 bezeichnen die 6 Szenarien im EAP-Test, die jeweils ein bis drei Gruppen von Items enthalten. Für Szenario 1 entfällt die Analyse, da es nur eine Itemgruppe umfasst.

### 3.3.4 Deskriptive Statistiken der verwendeten Variablen

Tabelle 3.4 können die deskriptiven Statistiken der in der Untersuchung verwendeten Messverfahren entnommen werden. Aufgrund des beschriebenen Problems, das beim Vorliegen von Testlet-Effekten die Reliabilitätsschätzungen überhöht sein können, werden sowohl für den EAP-Test als auch für den TBEU die Schätzungen für Cronbachs  $\alpha$  mit Testlets als Analyseeinheit berichtet. Für die Varianten mit 14 Itemgruppen, die sich auf die gleiche Emotionskontingenz beziehen, liegt die Schätzung für  $\alpha$  bei .74 und im Falle von 6 Szenarien als Analyseeinheit ist  $\alpha = .70$ .

Tabelle 3.4

*Deskriptive Statistiken und Reliabilitätsschätzungen für die verwendeten Testverfahren*

Test	Analyseeinheit	<i>N</i>	<i>M</i>	<i>SD</i>	$\alpha$
EAP-Test	6 (14) Testlets	210	0.61	0.11	.70 (.74)

Test	Analyseeinheit	<i>N</i>	<i>M</i>	<i>SD</i>	$\alpha$
TBEU	25 Testlets	149	0.70	0.10	.76
Me	10 Items	162	0.76	0.25	.80
An	20 Items	110	0.51	0.15	.58
Ma	20 Items	164	0.56	0.14	.59
Wi	28 Items	164	0.60	0.13	.58
O	9 Items	151	3.88	0.78	.75
C	9 Items	151	4.44	0.78	.82
E	9 Items	151	4.31	0.84	.83
A	9 Items	151	5.10	0.66	.85
N	9 Items	151	3.05	0.82	.78

*Anmerkungen.* Deskriptive Statistiken und Reliabilitätsschätzungen werden hier für den EAP-Test und TBEU nach ihrer Itemselektion berichtet.

### 3.3.5 Strukturprüfung des EAP-Tests und TBEU

Um die angenommene eindimensionale Struktur des EAP-Tests sowie des TBEU zu testen, wurden jeweils für beide Tests Konfirmatorische Faktorenanalysen (CFA) mit Mplus (Muthén & Muthén, 2010) durchgeführt.

#### *CFA des EAP-Tests*

Für den EAP-Test dienten die sechs Testlets, die jeweils die Items eines Szenarios umfassten, als Indikatoren. Zum Vergleich werden zudem die Ergebnisse einer CFA mit den 14 Itemgruppen als Indikatoren berichtet. Eine Inspektion der univariaten Verteilungen der Indikatoren deutete darauf hin, dass nicht von multivariater Normalverteilung der Indikatoren ausgegangen werden kann. Aus diesem Grund wurde als Schätzer Maximum Likelihood mit

Satorra-Bentler  $\chi^2$ -Teststatistik (MLM) verwendet, der robust gegenüber Verletzungen der Multivariaten Normalverteilung ist (Muthén & Muthén, 2010).

In Tabelle 3.5 werden Fit-Statistiken und die Verteilung der standardisierten Faktorladungen für den EAP-Test berichtet. Sowohl für die Variante mit 6 Indikatoren als auch für die Variante mit 14 Indikatoren kann der Modellfit als sehr gut bezeichnet werden (Hu & Bentler, 1999). Sowohl die globalen Fitstatistiken als auch die Ladungen sprechen dafür, dass ein eindimensionales Modell sehr gut auf die Daten passt.

Tabelle 3.5

*Fit-Statistiken und standardisierte Faktorladungen für die CFA des EAP-Tests*

Messmodell	Fit-Statistiken					Standardisierte Faktorladungen	
	$\chi^2$	<i>df</i>	<i>p</i>	CFI	RMSEA [90%-KI]	<i>M</i>	Range
EAP 6	7.26	9	.61	1.00	.00 [.00, .06]	.55	.28 - .64
EAP 14	96.72	77	.06	.94	.04 [.00, .06]	.59	.26 - .58

*Anmerkungen.* *N* = 210; KI = Konfidenzintervall.

***CFA des TBEU***

Die 25 Testlets des TBEU wurden in einer zufälligen Auswahl fünf Item-Parcels zugewiesen. Diese Parcels wurden als Indikatoren in der CFA verwendet, aufgrund der andernfalls hohen Anzahl von Indikatoren sowie des Umstands, dass einige der Testlets aus einzelnen dichotomen Items bestanden (für Vor- und Nachteile von Parcels als Indikatoren, siehe Little, Cunningham, Shahar & Widaman, 2002). Um Zufallseinflüsse zu berücksichtigen, wurde dieser Vorgang wiederholt und insgesamt drei Sets von Parcels gebildet, über die jeweils eine CFA durchgeführt wurde. Auch beim TBEU wiesen die univariaten Verteilungen darauf hin,



dass nicht von multivariater Normalverteilung ausgegangen werden kann. Als Schätzer wurde daher MLM verwendet.

In Tabelle 3.6 werden die Fit-Statistiken und standardisierten Faktorladungen für die drei Parcel-Varianten des TBEU berichtet. Die Ergebnisse sind zwischen den Parcel-Varianten stabil und sprechen insgesamt für eine sehr gute Passung des eindimensionalen Modells.

Tabelle 3.6

*Fit-Statistiken und standardisierte Faktorladungen für CFA des TBEU*

Messmodell	Fit-Statistiken					Standardisierte Faktorladungen	
	$\chi^2$	<i>df</i>	<i>p</i>	CFI	RMSEA [90%-KI]	<i>M</i>	Range
TBEU P1	9.42	5	.09	.98	.08 [.00, .15]	.65	.58 - .76
TBEU P2	2.38	5	.79	1.00	.00 [.00, .07]	.65	.52 - .77
TBEU P3	9.44	5	.09	.98	.08 [.00, .15]	.65	.57 - .78

*Anmerkungen.* *N* = 149; KI = Konfidenzintervall.

### 3.3.6 Faktorkorrelationen zwischen EU, Intelligenz und Persönlichkeitseigenschaften

Die Faktorkorrelationen zwischen dem EAP-Test, TBEU, Klassischer Intelligenz sowie den Big-Five Persönlichkeitsfaktoren wurden ebenfalls mit Mplus geschätzt. Als Indikatoren für den EAP-Test dienten die gebildeten 6 Testlets und für den TBEU die 25 Testlets. Die Items der Maße für klassische Intelligenz und den Big-Five wurden über Zufallsauswahlen zu Parcels zusammengefügt. Zum Vergleich der Stabilität der Ergebnisse werden in Anhang C bis D die Ergebnisse für zwei weitere Sets von zufällig gewählten Parcels berichtet.

Aufgrund der unterschiedlichen Designs, lagen nicht für alle Personen Daten auf allen Testverfahren vor. Daher wurden die Faktorkorrelationen in mehreren Modellen geschätzt. Als Schätzer wurde MLM verwendet.

### ***EAP-Test und Klassische Intelligenz***

Da für die Zusammenhänge des EAP-Tests mit Merkfähigkeit, Matrizen und Wissen aufgrund der unterschiedlichen Durchführungsdesigns mehr Datenpunkte vorlagen, als für die Zusammenhänge des EAP-Tests mit Analogien, wurden zwei Modelle geschätzt: a) ein Modell für die Faktorkorrelation zwischen EAP-Test und den Maßen für Standard-Intelligenz ohne Analogien und b) ein Modell auf Grundlage einer kleineren Teilstichprobe, bei der die Analogie-Aufgaben mit erhoben werden konnten. In Tabelle 3.7 werden die Faktorkorrelationen zwischen EAP-Test, Merkfähigkeit, Matrizen und Wissen präsentiert.

Tabelle 3.7

*EAP und Intelligenzmaße: Standardisierte Faktorladungen der Indikatoren und Faktorkorrelationen*

Indikatoren/ Faktoren	EAP	Me	Ma	Wi
EAP1_S	.32**			
EAP2_S	.63**			
EAP3_S	.61**			
EAP4_S	.61**			
EAP5_S	.65**			
EAP6_S	.54**			
ISTME1P1		.87**		
ISTME1P2		.72**		
ISTME1P3		.68**		

Indikatoren/ Faktoren	EAP	Me	Ma	Wi
ISTMA1P1			.52**	
ISTMA1P2			.60**	
ISTMA1P3			.53**	
ISTMA1P4			.48**	
ISTWI1P1				.72**
ISTWI1P2				.44**
ISTWI1P3				.34**
ISTWI1P4				.53**
Faktorkorrelationen				
Me	.26*			
Ma	.51**	.58**		
Wi	.26*	.29**	.43**	

*Anmerkungen.*  $N = 162$ ; Schätzer: MLM; Modell-Fit:  $\chi^2 = 102.45$ ,  $df = 113$ ,  $p = .75$ ; CFI = 1.00; RMSEA = .00, 90%-KI [.00, .03]; \* =  $p < .05$ , \*\* =  $p < .01$ .

Wie anhand der Tabelle zu erkennen ist, weist der EAP-Test mit .51 eine hohe positive Faktorkorrelation mit dem Matrizen-Test auf und jeweils mit .26 moderate positive Zusammenhänge mit Merkfähigkeit und Wissen. Betrachtet man die Zusammenhänge zwischen den Faktoren für klassische Intelligenz fällt auf, dass zwischen den Faktoren für Wissen und Matrizen eine moderate bis hohe Beziehung besteht.

In Tabelle 3.7 werden die Ergebnisse aus der kleineren Stichprobe für die Faktorkorrelationen zwischen EAP-Test und klassischer Intelligenz mit Analogien gezeigt. Wie aus der Tabelle entnommen werden kann, sind die Faktorkorrelationen zwischen dem EAP-Test und den Skalen für klassische Intelligenz nicht signifikant, mit Ausnahme einer

mittleren, positiven Faktorkorrelation zwischen EAP-Test und Matrizen. Unerwartet hoch ist die Faktorkorrelation zwischen Wissen und Analogien.

Tabelle 3.8

*EAP und Intelligenzmaße mit Analogien: Standardisierte Faktorladungen der Indikatoren und Faktorkorrelationen*

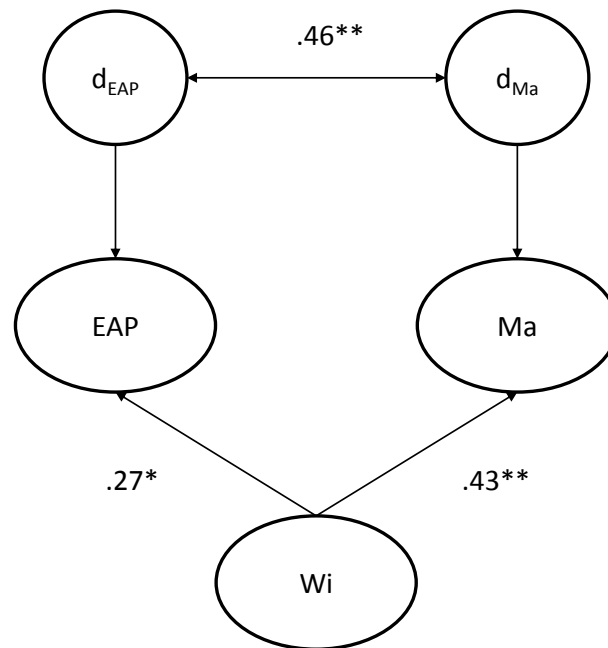
Indikatoren/ Faktoren	EAP	Me	An	Ma	Wi
EAP1_S	.15				
EAP2_S	.55**				
EAP3_S	.55**				
EAP4_S	.66**				
EAP5_S	.73**				
EAP6_S	.60**				
ISTME1P1		.85**			
ISTME1P2		.79**			
ISTME1P3		.71**			
ISTAN1P1			.46**		
ISTAN1P2			.52**		
ISTAN1P3			.32**		
ISTAN1P4			.68**		
ISTMA1P1				.53**	
ISTMA1P2				.57**	
ISTMA1P3				.26*	
ISTMA1P4				.41**	
ISTWI1P1					.64**

Indikatoren/ Faktoren	EAP	Me	An	Ma	Wi
ISTWI1P2					.38**
ISTWI1P3					.33**
ISTWI1P4					.54**
Faktorkorrelationen					
Me	.12				
An	.26	.54**			
Ma	.31*	.54**	.54**		
Wi	.09	.31**	.89**	.65**	

*Anmerkungen.*  $N = 108$ , Schätzer: MLM, Modell-Fit:  $\chi^2 = 183.21$ ,  $df = 179$ ,  $p = .40$ ; CFI = .99; RMSEA = .02, RMSEA 90%-KI [.00, .05]; \* =  $p < .05$ , \*\* =  $p < .01$ .

### ***Partielle Korrelation zwischen EAP-Test und Matrizen***

Da in dieser Untersuchung sowohl der EAP-Test als auch der Matrizen-Test (als Marker für Gf) substantielle Zusammenhänge mit dem Wissenstest (als Marker für Gc) aufwiesen, wurde zusätzlich die partielle Faktorkorrelation zwischen EAP-Test und Matrizen-Test unter Kontrolle von Wissen geschätzt. Die partielle Korrelation auf Faktorebene wurde als Strukturgleichungsmodell in MPlus mit MLM geschätzt. In Abbildung 3.1 wird das Strukturmodell gezeigt, mit der Schätzung für die partielle Korrelation sowie die standardisierten Pfadkoeffizienten.



*Abbildung 3.1.* Partielle Faktorkorrelation zwischen EAP-Test und Matrizen unter Kontrolle von Wissen. Illustriert wird das Strukturmodell. Als Indikatoren wurden die oben beschriebenen Testlets und Parcels verwendet.  $N = 164$ ; Schätzer: MLM; Modell-Fit:  $\chi^2 = 58.98$ ,  $df = 74$ ,  $p = .90$ ; CFI = 1.00; RMSEA = .00, RMSEA 90%-KI [.00, .02]; \* =  $p < .05$ , \*\* =  $p < .01$

### ***EAP-Test und Big-Five***

In Tabelle 3.9 werden die Faktorkorrelationen zwischen dem EAP-Test und den Big-Five Persönlichkeitseigenschaften gezeigt. Der EAP-Test weist eine mittlere positive Faktorkorrelation mit Offenheit für Erfahrung auf sowie eine kleine positive Faktorkorrelation mit Verträglichkeit. Mit den übrigen drei Persönlichkeitsfaktoren, Gewissenhaftigkeit, Extraversion und Neurotizismus, ist der EAP-Test unkorreliert.

Tabelle 3.9

*EAP und Big-Five Persönlichkeitsfaktoren: Standardisierte Faktorladungen der Indikatoren und Faktorkorrelationen*

Indikatoren/ Faktoren	EAP	O	C	E	A	N
EAP1_S	.20*					
EAP2_S	.57**					
EAP3_S	.57**					
EAP4_S	.66**					
EAP5_S	.72**					
EAP6_S	.63**					
OC_O_1P1		.76**				
OC_O_1P2		.65**				
OC_O_1P3		.84**				
OC_C_1P1			.75**			
OC_C_1P2			.97**			
OC_C_1P3			.57**			
OC_E_1P1				.80**		
OC_E_1P2				.83**		
OC_E_1P3				.80**		
OC_A_1P1					.77**	
OC_A_1P2					.81**	
OC_A_1P3					.90**	
OC_N_1P1						.74**
OC_N_1P2						.81**
OC_N_1P3						.73**

Indikatoren/ Faktoren	EAP	O	C	E	A	N
<b>Faktorkorrelationen</b>						
O	.35**					
C	.01	.02				
E	.00	.09	.10			
A	.16*	.28**	.29**	.52**		
N	.09	.19*	-.13	-.26**	-.12	

*Anmerkungen.*  $N = 152$ , Schätzer: MLM, Modell-Fit:  $\chi^2 = 196.31$ ,  $df = 174$ ,  $p = .12$ ; CFI = .98; RMSEA = .03, 90%-KI [.00, .05]; \* =  $p < .05$ , \*\* =  $p < .01$ .

### ***Faktorkorrelation zwischen EAP-Test und TBEU***

In Tabelle 3.10 werden die Schätzungen für das Modell der Faktorkorrelation zwischen EAP-Test und TBEU gezeigt. Mit einer Faktorkorrelation von .69 weist der EAP-Test, wie erwartet, die höchste Beziehung mit dem TBEU auf.

Tabelle 3.10

### ***EAP und TBEU: Standardisierte Faktorladungen der Indikatoren und Faktorkorrelation***

Indikatoren/ Faktoren	EAP	TBEU
EAP1_S	.15	
EAP2_S	.61**	
EAP3_S	.62**	
EAP4_S	.64**	
EAP5_S	.69**	
EAP6_S	.64**	
TBEU_1P1		.75**



Indikatoren/ Faktoren	EAP	TBEU
TBEU_1P2		.59**
TBEU_1P3		.64**
TBEU_1P4		.66**
TBEU_1P5		.62**
Faktorkorrelation		
TBEU	.69**	

Anmerkungen.  $N = 149$ , Modell-Fit:  $\chi^2 = 55.93$ ,  $df = 43$ ,  $p = .09$ ; CFI = .97; RMSEA = .05  
90%-KI [.00, .08]; \*\* =  $p < .01$ .

### *Faktorkorrelationen des TBEU*

Als ebenfalls neu entwickelter Test für EU, werden zudem auch für den TBEU die Beziehungen zu Intelligenz und Persönlichkeit im Sinne von konvergenter und diskriminanter Validitätsevidenz betrachtet. In Tabelle 3.11 werden die Faktorkorrelationen des TBEU mit Klassischer Intelligenz präsentiert. Der TBEU zeigt positive Zusammenhänge mit Merkfähigkeit und Wissen, die von mittlerer Höhe sind. Die Faktorkorrelation mit dem Matrizenest ist hingegen nicht signifikant. Die Ergebnisse sollten allerdings insgesamt vorsichtig betrachtet werden, da die Stichprobengröße für diese Zusammenhänge gering ist.

Tabelle 3.11

*TBEU und Intelligenz: Standardisierte Faktorladungen der Indikatoren und Faktorkorrelationen*

Indikatoren/ Faktoren	TBEU	Me	Ma	Wi
TBEU_1P1	.74**			
TBEU_1P2	.56**			

Indikatoren/ Faktoren	TBEU	Me	Ma	Wi
TBEU_1P3	.63**			
TBEU_1P4	.61**			
TBEU_1P5	.53**			
ISTME1P1		.87**		
ISTME1P2		.75**		
ISTME1P3		.72**		
ISTMA1P1			.55**	
ISTMA1P2			.56**	
ISTMA1P3			.22	
ISTMA1P4			.41**	
ISTWI1P1				.66**
ISTWI1P2				.31**
ISTWI1P3				.37**
ISTWI1P4				.59**
Faktorkorrelationen				
Me	.34**			
Ma	.23	.53**		
Wi	.30*	.30**	.65**	

*Anmerkungen.*  $N = 106$ , Schätzer: MLM, Modell-Fit:  $\chi^2 = 92.83$ ,  $df = 98$ ,  $p = .63$ ; CFI = 1.00; RMSEA = .00, 90%-KI [.00, .05]; \* =  $p < .05$ , \*\* =  $p < .01$ .

Aus Tabelle 3.12 können die Schätzungen des Modells für die Faktorkorrelationen zwischen TBEU und den Big-Five Persönlichkeitseigenschaften entnommen werden. Wie erwartet hat der TBEU eine moderate positive Beziehung mit Offenheit für Erfahrung und ist mit

Gewissenhaftigkeit, Verträglichkeit und Neurotizismus unkorreliert. Nicht erwartet wurde hingegen die niedrige negative Faktorkorrelation zwischen TBEU und Extraversion.

Tabelle 3.12

*TBEU und Big-Five Persönlichkeitsfaktoren: Standardisierte Faktorladungen der Indikatoren und Faktorkorrelationen*

Indikatoren/ Faktoren	TBEU	O	C	E	A	N
TBEU_1P1	.76**					
TBEU_1P2	.57**					
TBEU_1P3	.60**					
TBEU_1P4	.68**					
TBEU_1P5	.65**					
OC_O_1P1		.77**				
OC_O_1P2		.64**				
OC_O_1P3		.84**				
OC_C_1P1			.74**			
OC_C_1P2			1.00**			
OC_C_1P3			.54**			
OC_E_1P1				.81**		
OC_E_1P2				.82**		
OC_E_1P3				.80**		
OC_A_1P1					.74**	
OC_A_1P2					.80**	
OC_A_1P3					.88**	
OC_N_1P1						.74**
OC_N_1P2						.81**

Indikatoren/ Faktoren	TBEU	O	C	E	A	N
OC_N_1P3						.72**
Faktorkorrelationen						
O	.30**					
C	.04	.01				
E	-.18*	.11	.11			
A	-.03	.30**	.26**	.58**		
N	.07	.18*	-.13	-.24**	-.08	

*Anmerkungen.*  $N = 148$ , Schätzer: MLM, Modell-Fit:  $\chi^2 = 183.73$ ,  $df = 155$ ,  $p = .06$ ; CFI = .97; RMSEA = .04, 90%-KI [.00, .05]; \* =  $p < .05$ , \*\* =  $p < .01$ .

### 3.4 Diskussion

Insgesamt sprechen die Ergebnisse der ersten Studie für die psychometrische Qualität und Validität des EAP-Tests. Auf Ebene der Testlets liegen die Reliabilitätsschätzungen in einem akzeptablen Bereich, so dass der EAP-Test in seiner aktuellen Version in den folgenden Studien sinnvoll eingesetzt werden kann. Dennoch sollte eine Verbesserung der Reliabilität Ziel von Überarbeitungen des EAP-Tests sein, indem neue Items konstruiert werden, die die bisherigen, qualitativ guten Items ergänzen können. Dabei wäre es wünschenswert, dass gleichzeitig auch die Aspekte der Ökonomie und Zumutbarkeit in Betracht gezogen werden. Eine reine Erhöhung von Items erscheint wenig zielführend, da der EAP-Test in seiner jetzigen Form recht umfangreich und für die Testpersonen, durch den hohen Anteil zu lesender Texte, beanspruchend ist. Die bisherige Auswahl von Items enthält zudem Items von extremer Schwierigkeit bzw. Leichtigkeit, die Ziel zukünftiger Überarbeitungen des EAP-Tests sein sollten, da sie so wenig Information liefern können. Zwar wurde insgesamt das Ziel erreicht, eine Variation von Schwierigkeiten zu erzeugen und vor allem auch schwere Items

zu konstruieren, allerdings sollten in zukünftigen Studien die Merkmale der Items, die die im Rahmen des EAP die Schwierigkeit beeinflussen, genauer untersucht werden. In einer zusätzlichen Analyse der ausgewählten 63 Items des EAP-Tests waren Itemschwierigkeit und Anzahl der Emotionskontingenzen pro Zielperson unkorreliert ( $r = -.18$ ;  $p = .16$ ). Einschränkend muss für diese Studie erwähnt werden, dass für diesen ersten Entwurf von Items die Aspekte, von denen angenommen wurde, dass sie die Schwierigkeit beeinflussen, nicht systematisch variiert werden konnten. In zukünftigen Studien könnte dieser Aspekt fokussiert werden und über eine systematische Variation von Merkmalen untersucht werden.

Wie die Schätzungen für die Testlet-Effekte zeigen, sollte das Problem der Verletzungen der Unabhängigkeits-Annahme auch in zukünftigen Studien zum EAP-Test berücksichtigt werden. Wang et al. (2002) veranschaulichen die Bedeutung der Ausmaß der Testlet-Effekte, indem sie darauf aufmerksam machen, dass sie auf der gleichen Skala wie der Personenfähigkeiten-Parameter abgebildet sind. Testlet-Effekte größer 1 weisen darauf hin, dass der Anteil der Varianz, der auf den Testlet-Effekt zurückgeht, die Varianz zwischen den Personen übersteigt. Dabei kann nicht davon ausgegangen werden, dass das Problem von Testlet-Effekten auf den EAP-Test begrenzt ist und sollte daher im Rahmen ähnlicher Testverfahren in dieser Domäne mehr Aufmerksamkeit erhalten, wie es bisher nur in wenigen Studien erfolgte (Maul, 2012).

In dieser Studie konnte außerdem erste korrelative Validitätsevidenz für den EAP-Test gefunden werden. Die Erwartungen über die Beziehungen des EAP-Tests zu klassischer Intelligenz und den Big-Five Persönlichkeitseigenschaften wurden zum größten Teil bestätigt. Der EAP-Test weist ausschließlich positive Beziehungen zu Maßen der klassischen Intelligenz auf. Die Faktorkorrelationen liegen zudem in einer Höhe, die eine Interpretation als Form von Intelligenz erlaubt, ohne dass die, mit dem EAP-Test gemessene Fähigkeit untrennbar von etablierten Formen der Intelligenz ist. Wie erwartet weist der EAP-Test eine

höhere Beziehung zu Gf als zu Gc auf. Der beobachtete Zusammenhang zwischen EAP-Test und Gf bleibt bestehen, auch wenn das Maß für Gc, das auch mit Gf deutlich korreliert ist, in dieser Studie statistisch kontrolliert wird. Der Befund, dass der EAP-Test und das Maß für Gf deutlich korreliert sind, steht im Einklang mit der Definition von EU als Schlussfolgern über Emotionen sowie der Idee im EAP-Test, dass EU über das Erlernen neuer Emotionskontingenzen gemessen wird. Dennoch weisen bisherige Befunde zu den beiden verfügbaren Messinstrumenten, MSCEIT und STEU, ein anderes Bild auf, mit einer höheren Nähe zu Gc. In zukünftigen Studien sollte untersucht werden, ob das, was aktuell unter EU verstanden wird, vielleicht sinnvoll in Formen von Gf und Gc mit emotionalen Inhalten unterschieden werden könnte. Dass EU mit „Emotionswissen“ gleichzusetzen ist, wie es beispielsweise in der deutschen Version des MSCEIT der Fall (siehe Steinmayr et al., 2011), kann anhand dieser Ergebnisse allerdings angezweifelt werden.

Im Sinne diskriminanter Validitätsbelege ist die Faktorkorrelation des EAP-Tests mit Merkfähigkeit verbal in einer Höhe, die es sehr unwahrscheinlich macht, dass mit dem EAP-Test und seinem Aneignungs-Anwendungsphasen-Design lediglich das kurzfristige Behalten von Emotionsbegriffen gemessen wird. Trotzdem sollte berücksichtigt werden, dass sich dieses Ergebnis nur auf die hier erstellten Items beschränkt. Bei der Konstruktion von Tests basierend auf dem EAP sollte beachtet werden, dass das Testdesign oder bestimmte Merkmale von Items dazu führen könnten, dass Merkfähigkeit unter anderen Umständen einen großen Einfluss haben kann. Nichtsdestotrotz konnte gezeigt werden, dass eine Testkonstruktion mit dem EAP möglich ist, ohne dass verbale Merkfähigkeit einen großen Einfluss beim Lösungsprozess hat. In dem hier entwickelten Testdesign wurden einige Maßnahmen ergriffen, um den potentiellen Einfluss von interindividuellen Unterschieden in der Merkfähigkeit gering zu halten. Eine dieser Maßnahmen bezieht sich auf die visuelle Präsentation des EAP-Tests, für die unterschiedlich aussehende Avatare verwendet wurden,

die konsistent an bestimmten Orten des Bildschirms präsentiert werden. Eine Abgrenzung zu figuraler Merkfähigkeit wäre in zukünftigen Studien daher ebenfalls sinnvoll.

Ebenfalls im Sinne eines diskriminanten Validitätsbelegs kann das Muster an Beziehungen des EAP-Tests zu den Big-Five Persönlichkeitseigenschaften interpretiert werden. Der EAP-Test ist weitestgehend unabhängig von Persönlichkeitseigenschaften. Zudem weist der EAP-Test die zu erwartende positive Beziehung zu Offenheit für Erfahrung auf. Die hier gefundene geringe positive Beziehung zu Verträglichkeit, wurde hingegen nicht erwartet. Allerdings zeigen sich für die verfügbaren Verfahren, STEU und MSCEIT, ebenfalls immer wieder geringe bis mittlere positive Beziehungen zu Verträglichkeit (z. B. MacCann & Roberts, 2008; Rossen & Kranzler, 2009). Welche Gründe für diese konsistenten Zusammenhänge bestehen, ist bislang ungeklärt.

Im Sinne eines konvergenten Validitätsbelegs kann der hohe Zusammenhang zwischen EAP-Test und TBEU interpretiert werden. Anhand der Höhe der Faktorkorrelation von .69 kann allerdings davon ausgegangen werden, dass EAP-Test und TBEU nicht exakt das Gleiche messen. Bei der Interpretation des Zusammenhangs sollte auch berücksichtigt werden, dass für den TBEU insgesamt etwas schwächere Validitätsbelege hinsichtlich seiner Beziehung zu klassischer Intelligenz und Persönlichkeitseigenschaften gefunden wurden. Die, wenn auch geringe, negative Beziehung zu Extraversion ist nicht erwartungskonform. Zudem ist die vergleichsweise hohe Beziehung zu Merkfähigkeit schwierig zu deuten, auch wenn letzteres Ergebnis auf einer eher geringen Stichprobe beruht und vorsichtig interpretiert werden sollte. Insgesamt spricht die Tatsache, dass der EAP-Test eine vergleichsweise hohe Beziehung mit einem Test aufweist, der ebenfalls mit dem Ziel konstruiert wurde, EU zu erfassen, für die Validität des EAP-Tests.

Eine Limitation dieser Studie besteht in den verwendeten Messinstrumenten für klassische Intelligenz. Zum einen mussten in dieser Studie eine enge Auswahl von Maßen für

klassische Intelligenz getroffen werden, da die Testungsdauer begrenzt war. Zudem weisen die eingesetzten Maße zur Erfassung klassischer Intelligenz in dieser Stichprobe teilweise unerwartet geringe und nicht zufriedenstellende Reliabilitätsschätzungen auf (vergleiche Liepmann et al., 2006). In zukünftigen Studien wäre es daher sinnvoll, den Zusammenhang des EAP-Tests mit anderen Operationalisierungen für klassische Intelligenz zu untersuchen. Dies wäre auch vor dem Hintergrund sinnvoll, dass die hier gefundenen Ergebnisse repliziert werden müssen, um ausschließen zu können, dass die Ergebnisse allein auf Eigenschaften der hier eingesetzten Verfahren zurückzuführen sind.

In zukünftigen Studien sollte der EAP-Tests noch weiter in seinem nomologischen Netzwerk betrachtet werden. In diesem Zusammenhang könnten die Zusammenhänge des EAP-Tests mit den im MSCEIT operationalisierten Branches der EI sowie mit Messverfahren ähnlicher Konstrukte, wie Social Understanding, weitere korrelative Validitätsevidenz liefern. Trotz der genannten Limitationen dieser Studie und ausstehender Untersuchungen bezüglich der korrelativen Validitätsevidenz des EAP-Tests, sind die hier gefundenen Belege vielversprechend für den EAP-Test.



## 4. Studie 2

### 4.1 Hintergrund und Ziele

Im Rahmen des EAP wird die Annahme getroffen, dass die Testpersonen im Lösungsprozess auf Informationen aus der Aneignungsphase zurückgreifen. Die zweite Studie wurde mit dem Ziel durchgeführt, diese fundamentale Annahme über den Lösungsprozess im EAP zu testen. Die Annahme über den Lösungsprozess ist entscheidend, da mit ihr das Scoring sowie das Testdesign im EAP-Test begründet werden. Trifft die Annahme zu, dass Testpersonen Informationen über die Zielpersonen aus der Aneignungsphase zum Lösen der Aufgaben heranziehen, dann sollten sich Testantworten über eine Manipulation der Informationen über die Zielpersonen verändern lassen. Hat die Aneignungsphase hingegen keinen Einfluss auf den Lösungsprozess, sollte eine Manipulation der Aneignungsphase keine Auswirkung auf das Antwortverhalten der Getesteten haben. Eine Möglichkeit einer solchen Manipulation besteht darin, den Test um die Aneignungsphase zu kürzen, so dass der Testperson keine Informationen über die Emotionskontingenzen der Zielpersonen vorliegen. Werden diese Informationen der Testperson nicht vorgelegt, kann aus der Annahme, die dem EAP unterliegt, die Erwartung abgeleitet werden, dass sich die Testleistungen verschlechtern im Vergleich zu Testungen, bei denen die Aneignungsphase dargeboten wird. Diese Erwartung gilt unter der Voraussetzung, dass der mit dem EAP konstruierte Test Items enthält, für die Vorwissen über die Zielpersonen zur Lösung notwendig ist. Fehlen Informationen über die Zielpersonen, könnten die Testpersonen versuchen, die Items anhand von allgemeinem Wissen über die Regelmäßigkeit von Emotionen zu lösen oder zu raten (siehe Orchard et al., 2009). Eine Verschlechterung der Testleistung ist damit vor allem bei Items zu erwarten, die sich nicht ebenso gut durch allgemeines Wissen über die Regelmäßigkeit von Emotionen lösen ließen. Vor allem bei Items, in denen die Zielpersonen ungewöhnliche Emotionskontingenzen aufweisen, wird angenommen, dass diese ohne Vorwissen über die

Zielperson schwer vorhersagbar sind (Orchard et al., 2009). Wie in Studie 1 berichtet, wurde bei der Konstruktion des EAP-Tests versucht, auch Zielpersonen mit ungewöhnlichen Emotionskontingenzen zu erstellen, aufgrund der Annahme, dass dies die Schwierigkeit erhöhen kann. Sollten sich hingegen die Testantworten nach Wegfall der Aneignungsphase nicht verschlechtern, muss in Zweifel gezogen werden, dass die Aneignungsphase die angenommene Rolle während des Lösungsprozesses im EAP-Test spielt. Damit wäre auch fraglich, ob ein Test-Design und Scoring basierend auf dem EAP überhaupt sinnvoll ist.

Mit dem Ziel, die Annahme über die Rolle der Aneignungsphase im EAP zu testen, wurde in Studie 2 ein Experiment durchgeführt. Den Teilnehmerinnen und Teilnehmern wurde entweder der EAP-Test oder eine um die Aneignungsphase gekürzte Version des EAP-Tests vorgelegt, die im Folgenden als Application-Phase-Only Test (APO-Test) bezeichnet wird. Die Items des EAP- bzw. APO-Tests wurden vorab von Ratern in zwei Gruppen klassifiziert: a) Items, für deren Lösung ungewöhnliche Emotionskontingenzen eine Rolle spielen (d. h. vorab als schwer vorhersagbar eingeschätzte Items) oder b) Items, die gewöhnliche Emotionskontingenzen beinhalten (d. h. vorab als leicht vorhersagbar eingeschätzte Items). Es wurde erwartet, dass die Gruppe, die den APO-Test bearbeitet, schlechtere Leistungen aufweisen wird, als die Gruppe, die den EAP-Test bearbeitet, und dass schwächere Leistungen der APO-Gruppe besonders bei der Gruppe von Items zu beobachten sein werden, die vorab von den Ratern als schwer vorhersagbar eingeschätzt wurden.

## **4.2 Methode**

### **4.2.1 Stichprobe**

Die Analytestichprobe in Studie 2 umfasste insgesamt  $N = 158$  Personen. Zwei Personen waren zuvor aus der Stichprobe ausgeschlossen worden, da sie sich während der Testung wiederholt entgegen den Instruktionen verhielten und die Testung dieser Personen

infolgedessen abgebrochen werden musste. Die Stichprobe bestand aus Schülerinnen und Schülern zweier Gymnasien ( $n = 119$ ) sowie aus Personen, die in einer Kirchengemeinde akquiriert wurden ( $n = 39$ ). Unter der Bedingung ohne Aneignungsphase (d. h. APO-Test) wurden  $n = 79$  Personen (60.8 % weiblich) im Alter von 15 bis 29 Jahren, mit einem durchschnittlichen Alter von 18.81 Jahren ( $SD = 3.58$ ), getestet. Unter der Bedingung mit Aneignungsphase (d. h. EAP-Test) wurden ebenfalls  $n = 79$  Personen (63.3 % weiblich) getestet. Der Altersbereich der Mitglieder dieser Gruppe lag zwischen 15 und 30 Jahren, mit einem Mittelwert von 18.66 Jahren ( $SD = 3.38$ ).

#### 4.2.2 Messinstrumente

*EAP- und APO-Test.* Der in Studie 1 konstruierte EAP-Test wurde in der vollständigen, d. h. 76 Items umfassenden, Version durchgeführt. Der APO-Test wurde im Rahmen dieser Studie auf Grundlage des Testmaterials des EAP-Tests erstellt. Hierzu wurden sämtliche Abschnitte des EAP-Tests gestrichen, die zur Aneignungsphase zählen. Diese machten etwa ein Drittel bis die Hälfte des Testmaterials aus. Die Anwendungsphasen und Items wurden hingegen nicht verändert und sind somit identisch in EAP- und APO-Bedingung. Durch die Streichung der Aneignungsphasen fielen in einigen Szenarien auch Informationen weg, die für das Verständnis der beschriebenen Situationen wichtig erschienen (wie beispielsweise der Ort der Handlung), nicht aber über das vergangene emotionale Verhalten der Zielpersonen Auskunft geben. Fehlende Informationen dieser Art wurden in den Situationsbeschreibungen vor der Anwendungsphase ergänzt. Zudem wurde in allen Instruktionen der Hinweis entfernt, dass die Antworten auf Basis des Wissens über die Zielperson getroffen werden sollen.

*TBEU-Kurzversion.* Basierend auf vorläufigen Daten aus Studie 1, die zu diesem Zeitpunkt noch nicht abgeschlossen war, wurde eine Kurzform des TBEU erstellt. Für die Erstellung dieser Kurzversion wurden weichere Kriterien angesetzt, als es im Rahmen der finalen Itemselektion in Studie 1 der Fall war. Für die Selektion wurden Items mit Trennschärfen  $\leq 0$

ausgeschlossen. Weitere Selektionsmaßnahmen wurden vorerst nicht unternommen. Damit sollte für möglichst viele Items eine Entscheidung über Beibehaltung oder Verwerfung anhand der finalen Stichprobe ermöglicht werden, wobei gleichzeitig Items, die bereits zu diesem Zeitpunkt wenig vielversprechend erschienen, zugunsten einer kürzeren Durchführungsdauer ausgeschlossen werden konnten. In Einzelfällen wurden Items mit negativer Trennschärfe vorläufig beibehalten, wenn andernfalls durch den Wegfall der Items für eine Situation nur noch ein einzelnes Item verblieben wäre und sich somit die Struktur des Items stark verändert hätte. Die Kurzversion des TBEU umfasste 100 Items, die in 25 Itemgruppen genestet sind. Tabelle B.1, in Appendix B, können die Items entnommen werden, die für die Kurzversion ausgewählt wurden.

### **4.2.3 Experimentelles Design**

Den Testpersonen wurde in einem randomisierten Design als Between-Subject Faktor entweder der EAP-Test oder der APO-Test vorgelegt. Ein alternatives Vorgehen mit der Manipulation der Aneignungsphase als Messwiederholungsfaktor, in dem allen Personen sowohl EAP- als auch APO-Test vorgelegt werden, erschien nicht sinnvoll, da Erinnerungseffekte unvermeidbar gewesen wären. Die vorab eingeschätzte Vorhersagbarkeit der Items stellte in diesem Design einen Within-Subject Faktor dar. Die abhängige Variable war die Testleistung, die anhand des Auswertungsschlüssels des EAP-Tests ermittelt wurde.

### **4.2.4 Klassifikation der Items**

Die Klassifikation der Items wurde von zwei unabhängigen Ratern vorgenommen: einer Psychologin sowie einer Psychologie-Studentin, die im Rahmen ihrer Abschlussarbeit an der Untersuchung mitwirkte. Beiden Beurteilerinnen wurde der EAP-Test vorgelegt mit der Aufgabe, für jedes der 76 Items einzuschätzen, ob die im Item dargebotene emotionale Reaktion der Zielperson „ungewöhnlich“ oder „gewöhnlich“ ist. Die Items wurden anschließend als „leicht vorhersagbar“ klassifiziert, wenn beide Rater die emotionale

Reaktion als „gewöhnlich“ einschätzten und für diese Reaktion „wahrscheinlich“ die richtige Antwort war. Ebenfalls als „leicht vorhersagbar“ wurden Items klassifiziert, wenn die Rater eine Reaktion als „ungewöhnlich“ klassifizierten und „unwahrscheinlich“ die korrekte Antwort war. Entsprechend wurden Items als „schwer vorhersagbar“ klassifiziert, wenn die Rater die vorgelegte emotionale Reaktion als „gewöhnlich“ einschätzten und die korrekte Antwort „unwahrscheinlich“ war oder die emotionale Reaktion als „ungewöhnlich“ eingeschätzt wurde, diese aber „wahrscheinlich“ war. Während der Einschätzungen lagen den Ratern keine Informationen über die tatsächlichen, d. h. empirischen, Schwierigkeiten der Items vor.

Bei der Bestimmung des Grades der Übereinstimmung zwischen den Ratern wurde ein Cohen's  $\kappa$  von .44 ermittelt, das für das weitere Vorgehen als ausreichend beurteilt wurde. Für die Analysen wurden diejenigen Items verwendet, bei denen beide Rater zum gleichen Urteil gelangt waren. Insgesamt wurden 23 Items als leicht vorhersagbar und 31 Items als schwer vorhersagbar klassifiziert.

#### **4.2.5 Durchführung**

Die Schülerinnen und Schüler wurden in Klassenräumen ihrer jeweiligen Schulen getestet. Die Gemeindemitglieder wurden in einem Konferenzraum der Gemeinde getestet, der zu diesem Zweck zur Verfügung gestellt wurde. Die Teilnehmerinnen und Teilnehmer wurden in Gruppengrößen bis maximal 22 Personen von ein bis drei Testleiterinnen getestet. Die Instruktionen und der Ablauf der Testungen wurden in standardisierter Form durchgeführt. Die Teilnehmerinnen und Teilnehmer wurden randomisiert den beiden Gruppen zugewiesen, die entweder den EAP-Test oder APO-Test bearbeiteten. Die randomisierte Zuweisung wurde folgendermaßen vorgenommen: Die bei der Testung eingesetzten 22 Notebooks waren auf der Rückseite des Bildschirms mit Ziffern durchnummeriert. Die Ziffern stellten eine Kodierung dar, über die vorab festgelegt war, ob der EAP-Test oder der APO-Test durchzuführen ist. Die

Notebooks wurden von der Testleitung zufällig, d. h. ohne die Kodierung zu sehen, auf die Plätze im Testungsraum verteilt. Danach betraten die Teilnehmerinnen und Teilnehmer den Raum und verteilten sich auf die Plätze.

Zu Beginn der Testung wurden die demografischen Informationen mit einem Testheft erhoben. Anschließend bearbeiteten die Testpersonen die Kurzversion des TBEU im gleichen Testheft. Danach wurde der APO-Test oder der EAP-Test an den Notebooks durchgeführt. Im Anschluss wurden ein Leistungstest (die verbalen Wissensaufgaben aus dem I-S-T 2000 R) sowie ein Persönlichkeitstest (OCEANIC) dargeboten, die im Rahmen dieser Studie nicht von Relevanz sind. Die Durchführungsdauer variierte zwischen den Testpersonen, da der TBEU und EAP- bzw. APO-Test im selbstgewählten Tempo bearbeitet wurden.

#### **4.2.6 Auswertung**

Der EAP- und APO-Test wurden jeweils nach dem Auswertungsschlüssel des EAP-Tests bewertet, damit die Testleistungen miteinander vergleichbar sind. Nach dem Scoring wurden die Testleistungen bestimmt, indem jeder Person jeweils der Mittelwert a) über die Gruppe der als leicht vorhersagbar klassifizierten Items und b) über die Gruppe der als schwer vorhersagbar klassifizierten Items bestimmt wurde.

Um die vorhergesagten Effekte der Aneignungsphase auf die Testleistung für die als leicht bzw. schwer vorhersagbar klassifizierten Itemgruppen zu testen, wurde eine zweifaktorielle Varianzanalyse mit dem Between-Subjects Faktor Testbedingung (Faktorstufen: EAP vs. APO) sowie dem Within-Subjects Faktor eingeschätzte Vorhersagbarkeit (Faktorstufen: leicht vs. schwer) durchgeführt. Zur Prüfung der Homoskedastizitäts-Annahme für die ANOVA wurde der Levene-Test durchgeführt, dessen Ergebnis darauf hinwies, dass sich die Varianzen der Testcores in den Gruppen sowohl beim Subset der leicht vorhersagbaren Items ( $p = .33$ ) als auch bei den schwer vorhersagbaren Items ( $p = .97$ ) nicht signifikant voneinander unterscheiden. Eine Inspektion der Verteilungen

der Testwerte in den vier, aus dem 2×2-Design resultierenden, Gruppen ergab zudem, dass die Testwerte annähernd normalverteilt sind.

## **4.3 Ergebnisse**

### **4.3.1 Deskriptive Statistiken und Reliabilitätsschätzungen für die eingesetzten Testverfahren**

In Tabelle 4.2 werden deskriptive Statistiken der eingesetzten Testverfahren sowie Schätzungen für Cronbachs  $\alpha$  präsentiert. In dieser Studie wurden für den EAP-Test bzw. APO-Test keine Itemselektionen vorgenommen. Über beide Stichproben gepoolt erschien eine Itemselektion nicht sinnvoll, da die Trennschärfe- und Schwierigkeitsschätzungen als Konsequenz der Manipulation der Aneignungsphase in den Bedingungen unterschiedlich ausfallen könnten. Eine Inspektion der Daten bestätigte diese Vermutung. Aus diesem Grund erschien es ebenfalls nicht sinnvoll, die Item-Statistiken getrennt nach Gruppen zu betrachten und auf dieser Grundlage Items auszuwählen, die sowohl im APO als auch im EAP-Test ausgeschlossen werden. Mit diesem Vorgehen wäre eine zu große Anzahl von Items ausgeschlossen worden. Für den EAP-Test liegt die Schätzung für Cronbachs  $\alpha$  über 6 Szenarien als Testlets mit  $\alpha = .71$  in einer Höhe, die mit dem Ergebnis aus Studie 1 vergleichbar ist. Die Reliabilitätsschätzung für den APO-Test fiel mit  $\alpha = .50$  deutlich geringer aus.

Für die Kurzversion des TBEU wurde eine weitere Itemselektion anhand der Daten über beide Gruppen hinweg durchgeführt. Hierzu wurden Items mit Trennschärfen  $\leq .05$  ausgeschlossen. Die finale Version des TBEU bestand aus 86 Items, deren Trennschärfen von .07 bis .50 reichen, mit einer mittleren Trennschärfe von .26. Die Schätzung für Cronbachs  $\alpha$  über die 25 Itemgruppen als Analyseeinheiten liegt mit  $\alpha = .80$  in einer Höhe, die ebenfalls

vergleichbar mit dem Ergebnis aus Studie 1 ist. Insgesamt konnten in dieser Studie die Reliabilitätsschätzungen für EAP-Test und TBEU repliziert werden.

Tabelle 4.2

*Deskriptive Statistiken und Reliabilitätsschätzungen der in der Studie eingesetzten Verfahren*

Test	<i>N</i>	<i>M</i>	<i>SD</i>	Schiefe	Kurtosis	Anzahl Testlets	$\alpha$
APO	79	0.44	0.07	0.14	-0.84	6	.50
EAP	79	0.58	0.10	-0.33	0.14	6	.71
TBEU-K	158	0.72	0.11	-0.92	1.78	25	.80

*Anmerkungen:* TBEU-K = Kurzversion des TBEU. Die deskriptiven Statistiken und Testlets des EAP- und APO-Test liegt der vollständige Test mit 76 Items zugrunde. Für den TBEU-K werden alle Werte nach der Itemselektion berichtet. Für alle Tests wurde Cronbachs  $\alpha$  über ihre jeweiligen Testlets als Analyseeinheit geschätzt.

#### 4.3.2 Effekte auf die Testleistung

Die Ergebnisse der zweifaktoriellen Varianzanalyse für die Effekte der Aneignungsphase und Itemart auf die Testleistung werden in Tabelle 4.2 berichtet. Der Haupteffekt der Testbedingung ist signifikant und weist die erwartete Richtung auf. Testpersonen, die den Test unter der Bedingung mit Aneignungsphase bearbeiteten, lösten signifikant mehr Items ( $M = 0.58$ ,  $SD = 0.10$ ) als Testpersonen, denen die Aneignungsphase nicht vorgelegt wurde ( $M = 0.44$ ,  $SD = 0.07$ ). Ebenfalls signifikant und in erwarteter Richtung ist der Haupteffekt der vorab eingeschätzten Vorhersagbarkeit der Items auf die Testleistung. Die Testscores sind signifikant niedriger für das Subset von Items, das aufgrund ungewöhnlicher Emotionskontingenzen vorab als schwer vorhersagbar klassifiziert wurde ( $M = 0.40$ ,  $SD =$



0.14), als für die Items, die als leicht vorhersagbar klassifiziert wurden ( $M = 0.62$ ,  $SD = 0.12$ ). Außerdem ist die Interaktion zwischen Testbedingung und Vorhersagbarkeit signifikant. Wie erwartet ist die Minderung der Testleistung der APO-Gruppe besonders stark für Items, bei denen ungewöhnliche Emotionskontingenzen für die Lösung relevant sind. Die Richtung der Haupteffekte und Art der Interaktion wird in Abbildung 4.1 noch einmal illustriert.

Tabelle 4.2

*Effekt der Aneignungsphase und eingeschätzten Vorhersagbarkeit auf die Testleistung*

Quelle	<i>df</i>	<i>F</i>	$\eta$	<i>p</i>
Between subjects				
EAP vs. APO	1	68.42**	.55	< .01
Error	156			
Within subjects				
Vorhersagbarkeit	1	396.65**	.85	< .01
Vorhersagbarkeit $\times$ EAP vs. APO	1	25.81**	.38	< .01
Error	156			

*Anmerkungen.* Testleistung entspricht dem Mittelwert der Items nach Scoring mit dem Auswertungsschlüssel des EAP-Tests.

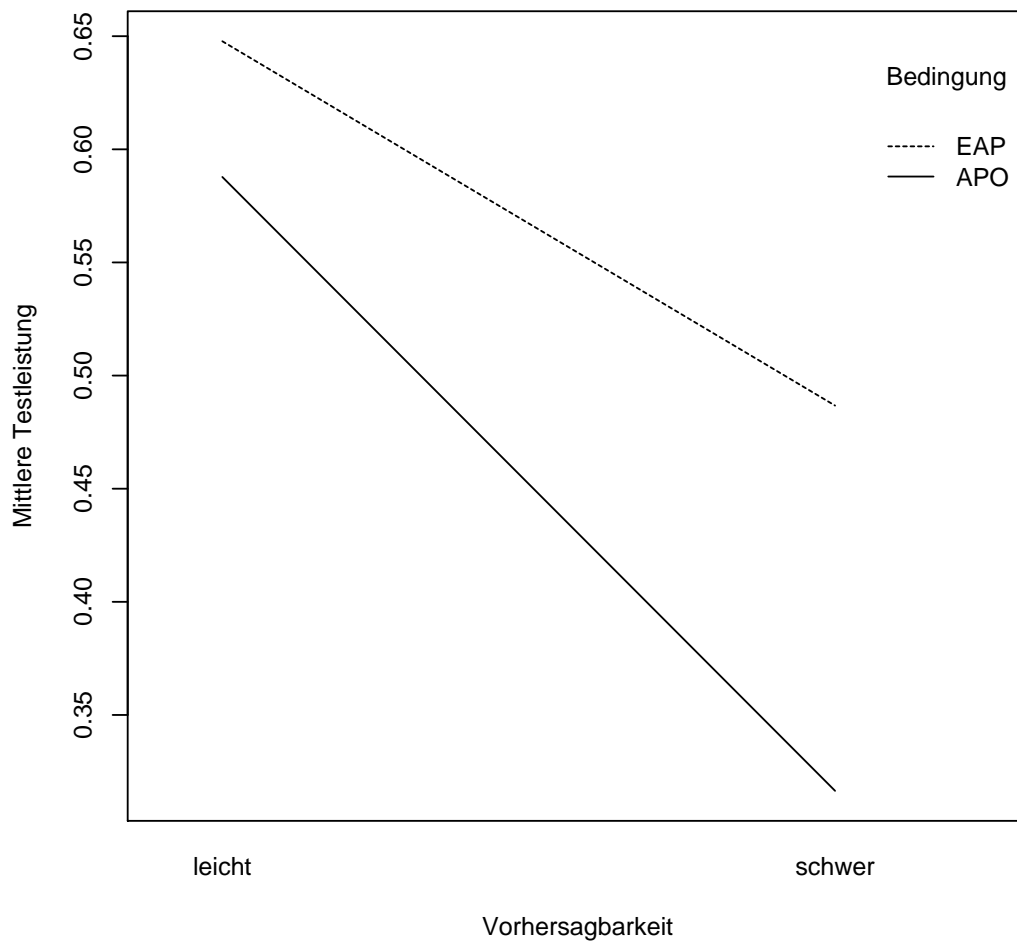


Abbildung 4.1. Interaktionsplot für den Effekt der Aneignungsphase und der vorab eingeschätzten Vorhersagbarkeit der Items auf die Testleistung.

#### 4.3.3 Kontrolle von Gruppenunterschieden hinsichtlich der EU-Fähigkeit

Um zu prüfen, ob die Randomisierung erfolgreich war und keine Unterschiede hinsichtlich der EU-Fähigkeit zwischen den Getesteten in der EAP- und APO-Gruppe bestanden, wurden die Testwerte des TBEU zwischen den beiden Gruppen verglichen. Zwischen den Testwerten der EAP-Gruppe ( $M = 0.73$ ,  $SD = 0.10$ ) und der APO-Gruppe ( $M = 0.72$ ,  $SD = 0.12$ ) im TBEU ist kein signifikanter Unterschied ( $t(156) = 0.88$ ,  $p = .38$ ) festzustellen. Dieses Ergebnis macht es äußerst unwahrscheinlich, dass die besseren Leistungen der Testpersonen

unter der EAP-Bedingung im Vergleich zur APO-Bedingung über vorab bestehende Unterschiede der Gruppen hinsichtlich ihrer EU-Fähigkeit erklärt werden kann.

#### 4.3.4 Zusammenhang der empirischen mit den vorhergesagten Itemschwierigkeiten

Im Rahmen der Konstruktion des EAP-Tests wurde vermutet, dass ungewöhnliche Emotionskontingenzen zu einer Erhöhung der Itemschwierigkeiten beitragen können. Der signifikante Haupteffekt der eingeschätzten Vorhersagbarkeit der Items auf die Testleistung weist darauf hin, dass ungewöhnliche Emotionskontingenzen zur Schwierigkeit von Items beitragen können.

In Tabelle 4.3 werden die Korrelationen zwischen den empirisch ermittelten Leichtigkeiten der Items und der vorab eingeschätzten Vorhersagbarkeit der Items gezeigt. Sowohl in der Gesamtstichprobe als auch in den Stichproben nach EAP- und APO-Test getrennt bestehen substantielle positive Zusammenhänge zwischen eingeschätzter Vorhersagbarkeit und der Leichtigkeit der Items. In der APO-Gruppe fällt dieser Zusammenhang am höchsten aus.

Tabelle 4.3

*Korrelationen zwischen empirischer Leichtigkeit und vorab eingeschätzter Vorhersagbarkeit*

Stichprobe	$r_{pb}$
Gesamt	.47**
APO-Bedingung	.57**
EAP-Bedingung	.30*

*Anmerkungen.* Den Analysen liegen die 54 Items zugrunde, die von den Ratern mehrheitlich als schwer oder leicht vorhersagbar klassifiziert wurden. Die vorab eingeschätzte Leichtigkeit ist als dichotome Variable mit 0 = schwer und 1 = leicht kodiert. \* =  $p < .05$ , \*\* =  $p < .01$ .

#### 4.4 Diskussion

Die Ergebnisse der experimentellen Untersuchung bestätigen, dass die Testpersonen beim Lösungsprozess im EAP-Test auf Informationen aus der Aneignungsphase zurückgreifen. Die experimentelle Manipulation der Aneignungsphase führte zum vorhergesagten Effekt auf die Testleistung. Wie erwartet erbrachten Testpersonen, denen die Aneignungsphase präsentiert wurde, bessere Testleistungen als Personen, die den Test ohne Aneignungsphase bearbeiteten. Ebenfalls den Erwartungen entsprechend, war die Testleistung bei den Items schlechter, die sich auf Zielpersonen mit ungewöhnlichen Emotionskontingenzen beziehen. Zudem wurde bestätigt, dass die Testleistung der APO-Gruppe bei Items mit ungewöhnlichen Emotionskontingenzen stärker gemindert war, was darauf hindeutet, dass für diese Items die Informationen aus der Aneignungsphase besonders relevant sind. Die Bestätigungen dieser Annahmen stellen einen wichtigen Validitätsbeleg für den EAP-Test dar, da der angenommene Lösungsprozess in diesen Ergebnissen widerspiegelt wird.

Die Ergebnisse liefern außerdem einen ersten Hinweis auf ein Merkmal von Items, das die Itemschwierigkeit im EAP-Test beeinflussen kann. Wie in Studie 1 beschrieben, wurden bei der Konstruktion des EAP-Tests Items erstellt, in denen die Zielpersonen ungewöhnliche emotionale Reaktionen aufweisen, da vorab vermutet wurde, dass dies die Schwierigkeit beeinflussen kann. Diese Vermutung wurde in dieser Studie bestätigt. In der APO-Bedingung verdeutlicht der hohe positive Zusammenhang zwischen Itemleichtigkeit und vorab eingeschätzter Vorhersagbarkeit, dass Vorwissen über die Zielpersonen besonders bei ungewöhnlichen Emotionskontingenzen notwendig ist. Das Merkmal der Ungewöhnlichkeit der emotionalen Reaktion senkt aber nicht nur die Lösungswahrscheinlichkeiten, wenn die Zielpersonen vorab nicht in einer vergleichbaren Situation beobachtet werden konnte, wie es im APO-Test der Fall ist. Wie der Haupteffekt der Vorhersagbarkeit auf die Testleistung sowie die Korrelation zwischen vorhergesagter und empirischer Itemschwierigkeit unter der

Bedingung mit Aneignungsphase verdeutlichen, tragen ungewöhnliche Emotionskontingenzen auch im EAP-Test zur Schwierigkeit der Items bei.

Die Ergebnisse geben auch einen Einblick in den Lösungsprozess von Personen, die nicht auf angeeignetes Wissen aus vorherigen Beobachtungen über die Zielperson zurückgreifen können. In der APO-Bedingung liegt die mittlere Lösungshäufigkeit für das Subset der Items, bei denen ungewöhnliche Emotionskontingenzen vorliegen, deutlich unter der Ratewahrscheinlichkeit. Ohne vorab erworbenes Wissen über die Zielpersonen, scheinen die Testpersonen also keinesfalls Zufallsantworten zu geben, sondern auf allgemeines Wissen zurückzugreifen und eher gewöhnliche emotionale Reaktionen für die Zielperson vorherzusagen – was bei diesen Items zu schlechteren Leistungen führt.

Eine Limitation dieser Studie besteht in der geringen Beurteilerübereinstimmung bei der Klassifikation der Vorhersagbarkeit der Items. Eine Schwierigkeit bei der Klassifikation der Items bestand darin, vorab genau zu spezifizieren, was unter „gewöhnlich“ verstanden wird. Die Klassifikationsergebnisse beruhen damit auf dem, was die Rater als gewöhnliche bzw. ungewöhnliche Reaktionen auffassen. Trotz der schwachen Übereinstimmung weisen die Ergebnisse darauf hin, dass die Klassifikation der Vorhersagbarkeit bei den verwendeten Items erfolgreich war.

Für die zukünftige Konstruktion von Tests basierend auf dem EAP wäre es hilfreich, vorab bestimmen zu können, ob eine Emotionskontingenz mehrheitlich als gewöhnlich oder ungewöhnlich wahrgenommen wird. Möglich wäre, dass emotionale Reaktionen dann als ungewöhnlich wahrgenommen werden, wenn sie z. B. nicht mit Vorhersagen aus Appraisal-Theorien (z. B. Roseman & Smith, 2001) übereinstimmen (Orchard et al., 2009), stereotypen Annahmen über das typische Verhalten von Mitgliedern sozialer Gruppen widersprechen oder auch, wenn sie nicht mit den eigenen typischen emotionalen Reaktionen übereinstimmen.

Letzteres wurde im Rahmen der Empathie-Forschung bereits als ein Faktor identifiziert, der die Perspektivübernahme erschweren kann (z. B. Lamm, Meltzoff & Decety, 2009).

Insgesamt liefert diese Studie experimentelle Validitätsevidenz für den EAP-Test, da die Ergebnisse die Annahmen über den Lösungsprozess im EAP-Test bestätigen. Die Ergebnisse unterstützen damit auch gleichzeitig das Scoring-Rational, das dem EAP zu Grunde liegt.

## 5. Studie 3

### 5.1 Hintergrund und Ziele

Mit der dritten Studie wurde das Ziel verfolgt, weitere korrelative Validitätsevidenz für den EAP-Test zu finden. In bisherigen Untersuchungen zur Validierung neuer Instrumente für EI und EU wurden Beziehungen zu Messansätzen ähnlicher oder möglicherweise identischer Konstrukte, die im Rahmen der Empathie- und Emotions-Forschung vorgeschlagen wurden, eher vernachlässigt. Dabei könnten Zusammenhänge mit Maßen für Theory of Mind, Affective Forecasting Ability oder Empathic Accuracy wichtige Validitätsbelege für einen neuen Test zur Erfassung von EU liefern.

Die wenigen bisher verfügbaren Studien zur Beziehung zwischen Messansätzen für EI und Konstrukten aus der Empathieforschung lieferten erste Belege für die theoretisch erwarteten Zusammenhänge. In einer Studie zum Zusammenhang zwischen Theory of Mind und EI haben Qualter et al. (2011) eine Jugendversion des MSCEIT sowie den Faux Pas-Test (Baron-Cohen, O'Riordan, Stone, Jones & Plaisted, 1999) einer Stichprobe von Kindern zwischen 8 und 10 Jahren dargeboten. Im Faux Pas-Test werden den Testpersonen Vignetten vorgelegt, die jeweils kurze Situationsbeschreibungen sowie Ausschnitte aus einem Gespräch zwischen zwei Personen beinhalten. In diesen Gesprächen löst eine der Zielpersonen unbeabsichtigt negative Emotionen in einer anderen Person aus. Die Aufgabe der Testpersonen ist es, diesen Fehltritt der Zielperson zu identifizieren und zu begründen, warum das Gesagte die Gefühle der betroffenen Person verletzt hat. In der Studie von Qualter et al. (2011) wurde der erwartete positive Zusammenhang zwischen den Testleistungen im MSCEIT und Faux Pas-Test ( $r = .38$ ) bestätigt, der selbst dann signifikant blieb ( $r_{ab.c} = .33$ ), wenn verbale Fähigkeiten und Alter statistisch kontrolliert wurden. Ein Zusammenhang vergleichbarer Höhe ( $r = .31$ ) wurde in einer Studie von Ferguson und Austin (2010)

zwischen EU, gemessen mit dem STEU, und dem Faux Pas-Test in einer Stichprobe mit Erwachsenen gefunden.

Der Zusammenhang von EI, erfasst mit dem MSCEIT, und Affective Forecasting Fähigkeit wurde in einer Studie von Dunn et al. (2007) untersucht. In dieser Studie wurde Affective Forecasting Fähigkeit über den Grad der Übereinstimmung von vorhergesagten Emotionen vor Eintreten eines Ereignisses (Ausgang einer Präsidentschaftswahl und Rückmeldung eines Prüfungsergebnisses) mit den selbstberichteten Emotionen nach Eintreten des Ereignisses operationalisiert. Mit dem MSCEIT konnten, den Erwartungen entsprechend, Fehler bei den Aufgaben zum Affective Forecasting vorhergesagt werden ( $\beta = -.22$ ) (Dunn et al., 2007). Orchard et al. (2009) schlugen vor, dass das Affective Forecasting Paradigma selbst einen nützlichen Ansatz zur Messung von EI liefern könnte oder als Variable für die Validierung eines EI-Tests dienen könnte. Des Weiteren wurde der Prozess des Affective Forecasting von Orchard et al. (2009) aus der Perspektive des EAP, das in diesem Beitrag noch als Emotion-based Principle Agent Paradigm bezeichnet wurde, betrachtet. Die Testpersonen könnten beim Affective Forecasting ihre Vorhersagen auf Grundlage der eigenen Ereignis-Emotions-Kontingenzen treffen (Orchard et al., 2009).

Studien zum Zusammenhang zwischen Empathic Accuracy und EU konnten zum Zeitpunkt dieser Studie interessanterweise nicht identifiziert werden. Aufgrund der hohen konzeptuellen Überlappung ist ein substantieller positiver Zusammenhang zwischen Maßen für EU und Empathic Accuracy zu erwarten. Die Fähigkeit, Emotionen bei anderen korrekt vorherzusagen zu können, ist sowohl Kern von Empathic Accuracy als auch ein wesentlicher Aspekt von EU. Hinsichtlich der Messung von Empathic Accuracy wurden von Ickes (2001) zwei Ansätze unterschieden: das Unstructured Dyadic Interaction Paradigma sowie das Standard Stimulus Paradigma.



Im Unstructured Dyadic Interaction Paradigma wird Empathic Accuracy gemessen, indem die Mitglieder einer Dyade die Emotionen und Kognitionen des jeweils anderen Mitglieds im Anschluss an eine Interaktion erschließen (siehe Ickes, 2001). Dazu werden die Dyaden verdeckt gefilmt, während sie in einer Laborsituation miteinander interagieren. Bei der Erfassung von Empathic Accuracy mit diesem Ansatz fungieren die Mitglieder der Dyade sowohl als Zielperson als auch als Testperson. Die Mitglieder der Dyade betrachten die Videos getrennt voneinander und berichten für selbst ausgewählte Stellen der Aufzeichnung ihre Emotionen und Kognitionen, die in diesem Moment der Interaktion bei ihnen aufgetreten sind. Anschließend betrachten die Mitglieder der Dyade das Video ein weiteres Mal unter der Instruktion, die Gedanken und Emotionen ihrer Interaktionspartnerin bzw. ihres Interaktionspartners an den ausgewählten Stellen zu erschließen. Die Messung von Empathic Accuracy ist in diesem Ansatz wenig standardisiert, da sich sowohl die Zielpersonen als auch der Verlauf der Interaktionen von Testperson zu Testperson unterscheidet. Im Standard Stimulus Paradigma wird Empathic Accuracy hingegen über standardisiertes Testmaterial erfasst, um einen Vergleich der Leistungen der Testpersonen zu ermöglichen (Ickes, 2001). Hierzu werden allen Testpersonen die gleichen Videoaufzeichnungen von Interaktionen unbekannter Zielpersonen vorgelegt, die zuvor für ausgewählte Stellen ihre Emotionen und Kognitionen berichtet haben. Auch hier besteht die Aufgabe der Testperson darin, anhand der Videoaufzeichnungen die Emotionen und Kognitionen der Zielpersonen in den ausgewählten Momenten der Interaktion zu erschließen.

Das EAP und die beiden Empathic Accuracy Paradigmen weisen Gemeinsamkeiten auf, unterscheiden sich aber auch in wesentlichen Gesichtspunkten voneinander. Gemeinsam ist den Ansätzen, dass die Testperson die Emotionen einer anderen Person korrekt vorhersagen soll. Die Grundlage, auf der die Testperson diese Vorhersage treffen soll bzw. kann, unterscheidet sich allerdings zwischen den drei Ansätzen. Im EAP greifen die Testpersonen bei der Vorhersage auf die Aneignungsphase zurück, mit der in standardisierter

Form Informationen über eine fiktive Zielperson dargeboten werden (siehe auch Studie 2). Im Rahmen des Unstructured Dyadic Interaction Paradigma ist die Zielperson der Testperson aufgrund früherer Interaktionen bekannt, so dass Informationen über die individuelle Regelmäßigkeiten von Emotionen der Zielperson vorliegen können. Die Aneignung dieser Informationen verläuft hierbei allerdings unstandardisiert, außerhalb der Testsituation. Analog zum Vorschlag von Orchard et al. (2009), den Prozess des Affective Forecastings aus der Perspektive des EAP zu betrachten, kann der Antwortprozess im Unstructured Dyadic Interaction Paradigma ebenfalls mit dem EAP in Verbindung gebracht werden. Die Testpersonen könnten ihre Vorhersagen über die Zielperson auf Ereignis-Emotions-Kontingenzen stützen, die sie in vorangegangenen Interaktionen mit der Zielperson beobachtet haben. Im Standard Stimulus Paradigma hingegen sind die Zielpersonen den Testpersonen unbekannt, so dass kein Wissen über individuelle Regelmäßigkeiten von Emotionen vorliegen können. Vorhersagen der Testpersonen könnten hier ausschließlich auf Basis von allgemeinen Regelmäßigkeiten von Emotionen getroffen werden.

Ein weiterer entscheidender Unterschied zwischen dem EAP und den Messansätzen für Empathic Accuracy besteht hinsichtlich des Scorings. Im Rahmen des EAP kann das Kriterium für das Scoring als veridikal betrachtet werden, da der Lösungsschlüssel im Schritt der Konstruktion der Aneignungsphase eindeutig festgelegt wird. Im Rahmen der Empathic Accuracy Paradigmen werden die selbstberichteten Emotionen der Zielpersonen als Kriterium für die Korrektheit der Antwort herangezogen, was der Methode des Target-Scoring entspricht.

Mit der dritten Studie sollte der Zusammenhang des EAP-Tests mit einem Messansatz aus der Empathie-Forschung untersucht werden. Angesichts der hohen konzeptuellen Überschneidung zwischen EU und Empathic Accuracy, erschien ein Instrument, bei dem auf den Empathic Accuracy Ansatz zurückgegriffen wird, als besonders geeignet. Aufgrund

mangelnder Verfügbarkeit von Instrumenten zur Erfassung von Empathic Accuracy, sollte in dieser Studie zunächst ein neuer Test entwickelt werden, der auf dem Empathic Accuracy Ansatz beruht. Allerdings wurde hierbei versucht, durch eine Abwandlung der von Ickes (2001) vorgeschlagenen Ansätze, die jeweiligen Vorteile des Standard Stimulus Paradigmas und des Unstructured Interaction Paradigmas miteinander zu vereinen. Vor diesem Hintergrund war das Ziel, einen SJT zu entwickeln, bei dem die Testperson für eine Reihe vorgegebener Situationen vorhersagen soll, wie die Partnerin bzw. der Partner emotional reagieren würde. Als Kriterium für die Richtigkeit der Vorhersage sollte hier die Vorhersage der Partnerin bzw. des Partners über die eigene Reaktion dienen, also die Affective Forecasts der Zielperson. Für den Prozess der Vorhersage von Affective Forecasts einer anderen Person haben Pollmann und Finkenauer (2009) den Begriff des Empathic Forecasting eingeführt. Durch die Vorgabe von Situationen wurde bezweckt, dass das Testmaterial standardisierter ist, als es unter dem Unstructured Interaction Paradigma der Fall ist, bei dem die Zielpersonen die Auswahl treffen. Gleichzeitig sollte durch die Vorgabe von bekannten Zielpersonen ermöglicht werden, dass die Testperson Wissen über die Zielperson durch vergangene Interaktionen erworben haben kann.

Der im Rahmen dieser Studie neu entwickelte SJT wurde als Dyadic Empathic Forecasting Test (DEF-Test) bezeichnet. Es wurde erwartet, dass die Leistungen im EAP-Test und die Genauigkeit der Vorhersagen über die Partnerin bzw. den Partner im DEF-Test substantiell positiv miteinander korreliert sind. Testpersonen mit hohen Werten im EAP-Test sollten die Emotionen ihrer Partnerin bzw. ihres Partners akkurater vorhersagen können, da angenommen wird, dass sie besser Wissen über emotionalen Regelmäßigkeiten dieser Personen aufbauen konnten und korrekt anwenden können. Aus dieser Perspektive, kann die Leistung im DEF-Test auch als Kriteriums-Messung für die Fähigkeit gesehen werden, die mit dem EAP-Test erfasst werden soll. Die Wahl geeigneter Kriterien zu Validierung eines

EI-Tests stellt eine wichtige Herausforderung dar, die bislang noch nicht zufriedenstellend gelöst werden konnte.

Bezüglich der Höhe des positiven Zusammenhangs zwischen EAP-Test und DEF-Test wird erwartet, dass der Erfolg der Aneignung von Wissen über die emotionalen Regelmäßigkeiten der Partnerin bzw. des Partners nicht ausschließlich von der Fähigkeit zum Erwerb und Anwendung von Emotionskontingenzen determiniert ist. Der Investment-Theorie entsprechend (siehe Cattell, 1963; von Stumm & Ackerman, 2013) ist zu erwarten, dass auch die Gelegenheit zum Erwerb eine Rolle spielen kann. Merkmale der Beziehung, wie beispielsweise die Dauer, Eigenschaften der Zielperson, wie die Expressivität von Emotionen (siehe z. B. Kring, Smith & Neale, 1994; Zaki et al., 2008), oder der Testperson, wie Offenheit für Gefühle (Ostendorf & Angleitner, 2004), könnten ebenfalls eine Rolle für den Erfolg für das Verständnis der Emotionen der Partnerin bzw. des Partners spielen. Aufgrund dieser Annahmen wird kein perfekter aber, aufgrund der konzeptuellen Nähe, ein nichtsdestotrotz starker positiver Zusammenhang zwischen den Testleistungen im EAP-Test und DEF-Test erwartet.

Bezüglich der möglichen Schlussfolgerungen über die Validität des EAP-Tests anhand dieser Studie, muss vorab einschränkend angemerkt werden, dass der EAP-Test hier anhand eines neuen Instrumentes erprobt wird, das selbst einer Validierung bedarf. Daher wird in dieser Studie auch Validitätsevidenz für den DEF-Test untersucht. Zum einen soll die angenommene eindimensionale Struktur des DEF-Tests getestet werden. Zum anderen werden die Beziehungen des DEF-Tests mit einem Maß für fluide, verbale Intelligenz und mit einem Instrument für die Big-Five Persönlichkeitsfaktoren untersucht. Aus den theoretischen Überlegungen und den Ergebnissen des EAP-Tests aus Studie 1 wird auch für den DEF-Test erwartet, dass dieser eine positive Beziehung zu verbalem Schlussfolgerndem Denken aufweist, der wiederum nicht so hoch ist, dass die Fähigkeiten untrennbar voneinander sind.

Darüber hinaus sollte der DEF-Test, ebenso wie der EAP-Test, ein Muster an Beziehungen zu den Big-Five Persönlichkeitsfaktoren aufweisen, wie es für einen Leistungstest erwartet werden kann: im Sinne eines diskriminanten Validitätsbelegs sollten, mit Ausnahme einer moderat positiven Beziehung zu Offenheit für Erfahrung, keine Beziehungen zu den Big-Five Persönlichkeitsfaktoren bestehen.

Bei der Untersuchung der Beziehung zwischen EAP-Test und DEF-Test sowie der Untersuchung weiterer Validitätsbelege für den DEF-Test, soll der Besonderheit der Datenstruktur Rechnung getragen werden. In dieser Studie werden daher Modelle angewendet, die für dyadische Datensätze angemessen sind.

## **5.2 Methode**

### **5.2.1 Entwicklung des DEF-Tests**

In dieser Studie wurde der DEF-Test zur Messung der Vorhersagegenauigkeit von Emotionen bei Partnerinnen und Partnern konstruiert. Der DEF-Test wurde als computergestütztes Verfahren entwickelt, das mit Hilfe der Programmiersoftware Inquisit 3 umgesetzt wurde. Im DEF-Test werden den Testpersonen 20 Situationen nacheinander vorgegeben, zu denen jeweils vier unterschiedliche Emotionsbegriffe vorgelegt werden. Die Aufgabe der Testperson besteht darin, für jede der vier Emotionen vorherzusagen, wie stark diese jeweils durch die beschriebene Situation ausgelöst werden würde. Die 20 Itemgruppen (d. h. je eine Situation gepaart mit 4 vorgegebenen Emotionen) werden den Testpersonen dabei in zwei Durchgängen präsentiert, wobei sich die Instruktionen in den beiden Durchgängen voneinander unterscheiden. Im ersten Durchgang besteht die Aufgabe der Testperson darin, für jede der vorgegebenen Emotionen vorherzusagen, wie stark ihre Partnerin oder ihr Partner diese Emotion in der jeweiligen Situation empfindet. Im zweiten Durchlauf soll die Testperson eine Vorhersage darüber treffen, wie stark die gegebene Situation die jeweiligen Emotionen in ihr

selbst auslösen würde. Die Situationen werden nacheinander auf dem Bildschirm präsentiert, zusammen mit den jeweils einzuschätzenden Emotionen. Diese vorgegebenen Emotionen können sich von Situation zu Situation unterscheiden. Die Vorhersage der emotionalen Reaktion für die Partnerin oder den Partner bzw. für sich selbst erfolgt, indem die Testperson einen Schieberegler auf einer 11-Punkte-Skala von 0 „gar nicht“ bis 10 „sehr stark“ bewegt. Der Schieberegler ist zu Beginn auf einem neutralen Punkt außerhalb der Skala positioniert. Die Testperson muss für jede Situation eine Einschätzung auf allen vier Emotionen abgeben, bevor sie zur nächsten Itemgruppe voranschreiten kann. Damit soll gewährleistet werden, dass die Testperson keine der zu treffenden Vorhersagen auslässt. In Abbildung 5.1 wird ein Bildschirmfoto einer Itemgruppe aus dem DEF-Test präsentiert. Das Beispiel ist dem ersten Durchgang des DEF-Tests entnommen, also unter der Instruktion sich vorzustellen, dass die Situation von der Partnerin bzw. dem Partner erlebt wurde.

Situation 8.  
Eine Person fährt mit ihrem Auto durch eine Stadt, in der sie noch nie war und sucht nach einer Straße. Als sie etwas langsamer wird, um das Straßenschild lesen zu können, hupt ein anderer Autofahrer hinter ihr und fährt dicht auf.

Stellen Sie sich vor, dass Ihre Partnerin/ Ihr Partner diese Person ist. Sagen Sie vorher, wie stark Ihre Partnerin/Ihr Partner die folgenden Emotionen fühlen würde:

38). nervös  
Ziehen Sie den Regler

gar nicht	1	2	3	4	5	6	7	8	9	sehr stark
-----------	---	---	---	---	---	---	---	---	---	------------

39). verlegen  
Ziehen Sie den Regler

gar nicht	1	2	3	4	5	6	7	8	9	sehr stark
-----------	---	---	---	---	---	---	---	---	---	------------

40). wütend  
Ziehen Sie den Regler

gar nicht	1	2	3	4	5	6	7	8	9	sehr stark
-----------	---	---	---	---	---	---	---	---	---	------------

41). erschrocken  
Ziehen Sie den Regler

gar nicht	1	2	3	4	5	6	7	8	9	sehr stark
-----------	---	---	---	---	---	---	---	---	---	------------

**weiter**

Abbildung 5.1. Screenshot einer Itemgruppe aus dem DEF-Test unter der Instruktion, die Vorhersage für die Partnerin bzw. den Partner zu treffen.

In den Situationsbeschreibungen wird die Zielperson, für die man eine Vorhersage treffen soll, mit dem Begriff „Person“ bezeichnet. Der Begriff „Person“ dient hierbei als Platzhalter und wird über die Instruktion genauer spezifiziert als entweder a) die Partnerin bzw. der Partner (im 1. Durchgang) oder b) die Testperson (im 2. Durchgang). Bei der Erstellung von Situationen und der Auswahl möglicher emotionaler Reaktionen wurde das Big-Five Persönlichkeitsmodell als loser Orientierungsrahmen herangezogen. Das Modell diente als Anhaltspunkt für die Konstruktion möglichst vielseitiger Situationen, die in Personen mit unterschiedlichen Ausprägungen auf den Big-Five verschiedene emotionale Reaktionen auslösen könnten. Die Erstellung von Items, bei denen interindividuelle Unterschiede hinsichtlich der wahrscheinlichen emotionalen Reaktion angenommen werden können, sollte sicherstellen, dass die Testpersonen bei der Vorhersage das Wissen über die individuelle Regelmäßigkeiten ihrer Partnerinnen bzw. Partner anwenden müssen, anstatt die Vorhersage allein auf Grundlage allgemeiner Regelmäßigkeiten stützen zu können, die für jede Zielperson mit hoher Wahrscheinlichkeit zutreffen würden.

### **5.2.2 Maße der Akkuratheit im DEF-Test**

Zur Bestimmung der Testleistung im DEF-Test wird die Höhe der Übereinstimmung zwischen den Vorhersagen der Testperson über die Zielperson und den Vorhersagen der Zielperson über sich selbst herangezogen. Um die Höhe der Übereinstimmung feststellen zu können, stehen prinzipiell mehrere Möglichkeiten zur Verfügung. Dabei lassen sich zwei Arten von Maßen unterscheiden, die hierfür in Frage kommen: Distanzmaße und korrelative Maße. Unter Verwendung dieser beiden Arten von Maßen werden zwei unterschiedliche Aspekte von Übereinstimmung betont. Während bei Distanzmaßen die Höhe der Abstände zwischen den Vorhersagen entscheidend ist, ist beim Einsatz korrelativer Maße der lineare Zusammenhang der Testantworten von Test- und Zielperson maßgebend. Da vorab kein Vorgehen aus inhaltlichen Gründen favorisiert wurde, erschien es sinnvoll, mehrere Varianten

zu prüfen und diejenigen Maße im weiteren Verlauf zu verwenden, die für den DEF-Test die höchste psychometrische Qualität liefern. Als Distanzmaß wurde zum einen die Summe quadrierter Differenzen zwischen den Vorhersagen herangezogen, was laut Cronbach (1955) der klassischen Art des Vorgehens bei der Bestimmung der Akkuratheit entspricht. Als weiteres Distanzmaß wurde die euklidische Distanz ermittelt, d. h. die positive Quadratwurzel aus der Summe quadrierter Differenzen zwischen den Vorhersagen. Als korrelatives Maß zur Bestimmung der Akkuratheit wurde Spearmans Rangkorrelation zwischen den Vorhersagen der Test- und Zielperson bestimmt. Dabei wurden zwei Varianten erprobt: Bei der ersten Variante, wurde die Rangkorrelation innerhalb jeder Itemgruppe bestimmt. Dies bedeutet, dass die Rangkorrelationen über je 4 Items gebildet wurden und somit 20 Rangkorrelationen pro Person als Indikatoren für die Akkuratheit zur Verfügung standen. Bei der zweiten Variante wurden die Rangkorrelationskoeffizienten über eine größere Anzahl von Items bestimmt. Hierbei wurde die Rangkorrelation über 16 Items ermittelt, womit für jede Testperson 5 Rangkorrelationskoeffizienten als Indikatoren der Leistung bestimmt wurden. Diese Variante wurde erprobt, da so jedem Rangkorrelationskoeffizienten eine größere Anzahl an Datenpunkten zu Grunde liegt, was die Genauigkeit der Schätzungen der einzelnen Rangkorrelationen verbessern sollte.

### **5.2.3 Analysen dyadischer Daten**

Für die Schätzung des Zusammenhangs zwischen den Testleistungen im EAP- und DEF-Test wurden Dyaden getestet, wobei jedes Dyadenmitglied einmal als Testperson in die Stichprobe einfluss und einmal als Zielperson das Kriterium für das Scoring im DEF-Test bot. Für dyadische Datensätze muss angenommen werden, dass die Datenpunkte nicht unabhängig voneinander sind (Kenny, Kashy & Cook, 2006).

Im Rahmen der Analyse dyadischer Daten wurden verschiedene Modelle vorgeschlagen, mit denen die Abhängigkeiten der Datenpunkte berücksichtigt werden können



(siehe z. B. Kenny et al., 2006; Olsen & Kenny, 2006). Für die Untersuchung der Beziehung zwischen DEF-Test und EAP-Test wurde in dieser Studie das Actor-Partner Interdependence Modell (APIM) eingesetzt. Zur Prüfung der eindimensionalen Struktur des DEF-Tests wurde zudem eine dyadische CFA durchgeführt.

Bei der Wahl geeigneter Modelle bezieht sich eine weitere wichtige Frage darauf, ob die Mitglieder der Dyade sinnvoll unterschieden werden können (siehe Kenny et al., 2006). Unterscheidbarkeit ist dann gegeben, wenn die Mitglieder der Dyade hinsichtlich einer bedeutsamen Variablen zwei Gruppen zugeteilt werden können (Kenny et al., 2006). Beispiele hierfür sind Untersuchungen, in denen sich die Dyaden aus einem Elternteil und Kind oder einem Vorgesetzten und einem Mitarbeiter zusammensetzen (Kenny et al., 2006). Obwohl in vielen Untersuchungen mit Paaren die Variable Geschlecht ein inhaltlich bedeutsames Unterscheidungsmerkmal der Dyaden-Mitglieder darstellt, spielt dies im Rahmen der vorliegenden Fragestellung keine Rolle. Die Dyaden-Mitglieder werden in dieser Untersuchung vielmehr als ununterscheidbar betrachtet. Olsen und Kenny (2006) haben sowohl für das APIM als auch für die dyadische CFA bestimmte Parameter-Restriktionen sowie eine Adjustierung der Fit-Statistiken vorgeschlagen, so dass sich die Modelle auf Dyaden mit ununterscheidbaren Mitgliedern sinnvoll anwenden lassen.

### ***Dyadische Datenstruktur***

Um das APIM und die dyadische CFA schätzen zu können, müssen die Daten vorab in eine bestimmte Datenstruktur überführt werden. Eine Möglichkeit der Anordnung der Daten besteht in der dyadische Datenstruktur (siehe Kenny et al., 2006). Bei der dyadischen Datenstruktur stellen die Dyaden die einzelnen Datenpunkte dar. Für jede Dyade liegen  $2 \times k$  Variablen vor, wobei  $k$  die Anzahl der Variablen ist, die für jedes Dyaden-Mitglied gemessen wurde (Kenny et al., 2006). Diese Form der Anordnung der Daten macht eine Zuweisung der Dyaden-Mitglieder zu einem der beiden Sets von  $k$  Variablen notwendig. Da es sich in dieser

Untersuchung um ununterscheidbare Dyaden-Mitglieder handelt, wurden die Partnerinnen bzw. Partner einer Dyade willkürlich einem der beiden Sets von Variablen zugeordnet. Im Folgenden werden diese beiden Sets von Variablen mit P1 und P2 bezeichnet, was für Partner/in 1 bzw. Partner/in 2 steht.

### ***APIM***

Das APIM ist ein Regressionsmodell für dyadische Daten, mit dem die Abhängigkeiten der Mitglieder der Dyaden modelliert werden können. In Abbildung 5.2 wird das APIM illustriert. In dieses Modell gehen die Daten der Individuen ein, wobei die Individuen genestet in ihren Dyaden behandelt werden (Cook & Kenny, 2005). Das Modell beinhaltet eine Prädiktor- und eine Kriteriums-Variable, die jeweils für beide Mitglieder der Dyade gemessen wurden (Kenny et al., 2006). Im APIM werden je zwei sogenannte Actor- und Partner-Effekte voneinander unterschieden, die eine zentrale Rolle für die vorliegende Fragestellung spielen: Actor-Effekte stellen die regressive Beziehung zwischen einer Prädiktor- und einer Kriteriums-Variablen innerhalb eines Individuums dar. Die Partner-Effekte hingegen entsprechen der Regression der Kriteriums-Variable eines Individuums auf die Prädiktor-Variable seiner Partnerin bzw. seines Partner. Ein weiteres Merkmal des APIM besteht darin, dass die Korrelation zwischen den Prädiktor-Variablen der Mitglieder der Dyaden frei geschätzt wird. Hierdurch werden bei der Schätzung der Actor-Effekte die Partner-Effekte kontrolliert und umgekehrt (Cook & Kenny, 2005). Außerdem wird die Korrelation zwischen den Residuen der Partnerinnen und Partner frei geschätzt, womit weitere Quellen für die Abhängigkeit der Dyaden berücksichtigt werden, die nicht bereits im APIM durch die Prädiktor-Variablen modelliert werden (Cook & Kenny, 2005; Kenny et al., 2006).

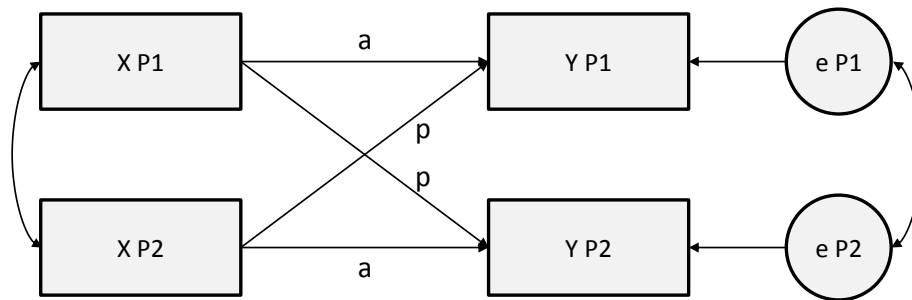


Abbildung 5.2. Beispielhafte Darstellung eines APIM mit einer Prädiktor-Variable X und einer Kriteriums-Variable Y für Partner/in 1 (P1) und Partner/in 2 (P2). a = Actor-Effekt; p = Partner-Effekt.

Olsen und Kenny (2006) haben ein Set von Restriktionen vorgeschlagen, mit denen das APIM sinnvoll für den Fall ununterscheidbarer Mitglieder der Dyaden eingesetzt werden kann. Hierzu werden nach Olsen und Kenny (2006) folgende Parameter für die Partnerin bzw. den Partner 1 und die Partnerin bzw. den Partner 2 gleichgesetzt: die Actor-Effekte, die Partner-Effekte, die Mittelwerte der Prädiktoren, die Varianzen der Prädiktoren, die Interzepte der Kriteriums-Variable sowie die Residualvarianzen.

### ***Dyadische CFA***

Olsen und Kenny (2006) haben zudem ein Modell für eine dyadische CFA vorgeschlagen, das für die Anwendung bei ununterscheidbaren Dyaden-Mitgliedern geeignet ist. In einer dyadischen CFA wird jede latente Variable sowohl für die Partnerin bzw. den Partner 1 als auch die Partnerin bzw. den Partner 2 spezifiziert, wobei die Korrelationen der entsprechenden Faktoren frei geschätzt werden. Zudem werden die Korrelationen zwischen den Residuen paarweise frei geschätzt. Für den Fall ununterscheidbarer Dyaden schlugen Olsen und Kenny (2006) vor, die folgenden Parameter jeweils paarweise gleichzusetzen: die

Varianzen der entsprechenden Faktoren, die Ladungen und Interzepte der entsprechenden manifesten Variablen sowie die Varianzen der entsprechenden Residuen.

### ***Adjustierung des Modell-Fits***

Kenny et al. (2006) wiesen darauf hin, dass eine arbiträre Zuweisung der Daten der Dyaden-Mitglieder zu den beiden Sets von  $k$  Variablen einen zusätzlichen und irrelevanten Einfluss auf den Modellfit haben kann. Olsen und Kenny (2006) schlugen daher eine Anpassung des  $\chi^2$ -Wertes, der Freiheitsgrade sowie der Fit-Indizes vor. Im Rahmen dieser Adjustierung wird zunächst ein angepasstes saturiertes Modell spezifiziert, indem bestimmte Restriktionen eingeführt wurden (siehe Kenny et al., 2006). Dieses Modell wird als interchangeable saturated Modell (I-SAT) bezeichnet (Olsen & Kenny, 2006). Der  $\chi^2$ -Wert und die Freiheitsgrade des interessierenden Modells werden anhand des  $\chi^2$ -Wertes und der Freiheitsgrade des I-SAT Modells korrigiert. Zur Bestimmung einiger Fit-Indizes wird zudem ein ebenfalls über bestimmte Restriktionen angepasstes Null-Modell als neues Referenzmodell spezifiziert, was als interchangeable Null-Modell (I-NULL) bezeichnet wird.

### **5.2.4 Stichprobe**

Insgesamt wurden 146, überwiegend heterosexuelle (99.3 %), Paare getestet. 50.3 % der Getesteten waren weiblich. Das Alter der Teilnehmerinnen und -teilnehmer lag zwischen 16 und 75 Jahren mit einem durchschnittlichen Alter von 31.2 Jahren ( $SD = 12.67$  Jahre). Die Beziehungsdauer reichte von 1 Monat bis 51 Jahren, wobei die mittlere Beziehungsdauer bei 8.4 Jahren ( $SD = 11.62$  Jahre) lag. 93.5% der Getesteten gaben an, Deutsch als Muttersprache zu haben. Alle Personen, die nicht Deutsch als Muttersprache hatten, sprachen seit mindestens 5 Jahren Deutsch. Die Paare wurden unter den Psychologiestudierenden der Bergischen Universität Wuppertal angeworben sowie im privaten Umfeld von 5 Studierenden, die im Rahmen ihrer Abschlussarbeiten an den Erhebungen mitwirkten.

### 5.2.5 Messinstrumente

*EAP-Test.* Der in Studie 1 entwickelte EAP-Test wurde in der vollständigen (d. h. 76 Items umfassenden) Version eingesetzt.

*DEF-Test.* Der im Rahmen dieser Studie beschriebene DEF-Test wurde den Testpersonen vorgelegt. Die 20 Itemgruppen wurden in einem ersten Durchgang unter der Instruktion der Partnereinschätzung vorgelegt und anschließend in einem zweiten Durchgang unter der Instruktion, die emotionalen Reaktionen für sich selbst einzuschätzen.

*Berliner Test zur Erfassung fluider und kristalliner Intelligenz 11+ – Schlussfolgerndes Denken – Verbaler Teil.* Verbale fluide Intelligenz wurde mit dem Untertest Schlussfolgerndes Denken – verbaler Teil aus dem Berliner Test zur Erfassung fluider und kristalliner Intelligenz (BEFKI; Wilhelm, Schroeders & Schipolowski, 2014) erfasst. Beim BEFKI handelt es sich um eine Serie von Tests, die jeweils für verschiedene Altersklassen geeignet sind. In dieser Studie wurde die Variante 11+ eingesetzt, die für Schülerinnen und Schüler der gymnasialen Oberstufe und den Einsatz bei Erwachsenen geeignet ist. Verbales schlussfolgerndes Denken wird im BEFKI über 16 Items operationalisiert. Innerhalb jedes Items werden den Testpersonen mehrere Prämissen vorgegeben, zu denen eine Frage gestellt wird, deren korrekte Antwort sich auf Grundlage der vorgegebenen Prämissen logisch erschließen lässt. Zur Beantwortung der Fragen werden 3 bis 4 Optionen vorgelegt, aus denen die Testperson die logisch Korrekte wählen soll. Der BEFKI ist als Paper-&-Pencil-Test entwickelt worden. Im Rahmen dieser Studie wurde für den eingesetzten Untertest schlussfolgerndes Denken – verbal eine computergestützte Version mit Hilfe von Inquisit 3 erstellt und bei den Testungen eingesetzt. Für die Bearbeitung des Untertests ist im Manual des BEFKI eine zeitliche Begrenzung von 14 Minuten vorgegeben. Die computergestützte Variante wurde so programmiert, dass die Testung nach Erreichen dieser Zeitbegrenzung

automatisch abbricht. Eine computergestützte Version erschien im Hinblick auf die Durchführung des Tests als besonders praktikabel.

*24-Adjective Measure (24-AM)*. Als Marker für die Big-Five Persönlichkeitsfaktoren wurde das von Herzberg und Brähler (2006) entwickelte 24-Adjective Measure (24-AM) eingesetzt. Im 24-AM werden den Testpersonen Adjektive vorgegeben, für die sie jeweils auf einer Skala von 1 „starke Ablehnung“ bis 7 „starke Zustimmung“ angeben sollen, wie sehr diese Persönlichkeitseigenschaft auf sie zutrifft. Als Grundlage für die Erstellung der Kurzsкала wurde von Herzberg und Brähler (2006) zunächst eine deutsche Übersetzung des Ten-Item Personality Inventory (TIPI; Gosling, Rentfrow & Swann, 2003) angefertigt, die die Autoren anschließend um neue Items ergänzten, mit dem Ziel der Verbesserung der psychometrischen Qualität des Instrumentes. Die finale Version umfasst 24 Adjektive, die in Anbetracht der Kürze des Verfahrens zufriedenstellende psychometrische Eigenschaften besitzt (Herzberg & Brähler, 2006). Das 24-AM wurde ebenfalls in Form einer Computerversion durchgeführt.

### **5.2.6 Durchführung**

Die Paare wurden teilweise in Räumen der Bergischen Universität Wuppertal und teilweise in ihrem häuslichen Umfeld getestet. Die Anwerbungen und Erhebungen fanden in Kooperation mit fünf Psychologiestudierenden statt, die ihre Abschlussarbeiten im Zeitraum von Sommer 2012 bis Frühjahr 2013 durchführten. Die Untersuchung wurde mit zwei verschiedenen Designs durchgeführt: Bei den ersten 75 Paaren wurde zunächst der EAP-Test, danach der demografische Fragebogen, der DEF-Test und abschließend der 24-AM vorgelegt. Bei den darauf folgenden 71 Paaren wurde zuerst der demografische Fragebogen, der DEF-Test, der BEFKI 11+ und abschließend der EAP-Test durchgeführt.

### **5.2.7 Auswertung**

Als Grundlage für die Analysen wurde die dyadische Struktur als Methode der Anordnung der Daten gewählt (siehe z. B. Kenny et al., 2006). Für das APIM und die dyadische CFA wurden die von Olsen und Kenny (2006) vorgeschlagenen Restriktionen für den Fall ununterscheidbare Dyaden-Mitglieder gesetzt sowie die Adjustierung der Fitstatistiken und Freiheitsgrade durchgeführt (siehe auch Peugh, DiLillo & Panuzio; 2013). Die Modelle wurden mit AMOS geschätzt, mit Maximum Likelihood (ML) als Schätzer.

Aufgrund der, durch die Durchführungs-Designs bedingten, relativ geringen Stichprobenumfänge, die für die Beziehungen mit verbalem schlussfolgernden Denken und den Big-Five Persönlichkeitsfaktoren zur Verfügung standen, wurden diese Beziehungen, anders als in Studie 1, auf Ebene manifester Variablen geschätzt. Zudem werden diese Korrelationen getrennt für die Zufallsauswahlen für Partnerin bzw. Partner 1 und 2 berichtet.

## **5.3 Ergebnisse**

### **5.3.1 Itemanalysen des EAP-Tests und DEF-Tests**

Zunächst wurden Itemanalysen sowohl für den EAP-Test als auch für den DEF-Test durchgeführt. Tabelle 5.1 können die Ergebnisse der Itemanalysen entnommen werden. Auf Grundlage der vorliegenden Stichprobe wurde für den EAP-Test eine Itemselektion durchgeführt. Hierzu wurden sämtliche Items mit Trennschärfen  $\leq .05$  ausgeschlossen. Wie in Studie 1 wurde die Itemselektion des EAP-Test auf Ebene der Items anstatt auf Ebene der Testlets vorgenommen, um ungeeignete Items innerhalb eines Testlets zu identifizieren, ohne potentiell gute Items zu verwerfen. In Tabelle 5.1 sind die Itemstatistiken für den EAP-Test vor und nach der Itemselektion gegenübergestellt. Nach der Itemselektion verblieben für den EAP-Test 61 Items, deren Schätzungen für die Trennschärfen zwischen .05 und .48 lagen, mit einer mittleren Trennschärfe von .22.

Tabelle 5.1

*Zusammenfassung der Itemmittelwerte und Trennschärfen für den EAP-Test und DEF-Test*

Test	Anzahl Items	Itemmittelwerte			Trennschärfen		
		<i>M</i>	<i>SD</i>	Range	<i>M</i>	<i>SD</i>	Range
EAP-Test							
unselektiert	76	0.62	0.25	0.07 – 0.97	.16	.12	-.15 – .47
selektiert	61	0.64	0.24	0.07 – 0.96	.22	.10	.05 – .48
DEF-Test							
EukDis	20	5.93	0.61	4.47 – 6.85	.27	.10	.08 – .42
QuaDiff	20	43.33	7.50	27.47 – 56.45	.17	.05	.05 – .27
$r_s$ 4 Items	20	0.41	0.19	0.07 – 0.76	.05	.08	-.13 – .15
$r_s$ 16 Items	5	0.46	0.10	0.30 – 0.56	.16	.08	.07 – .25

*Anmerkungen.*  $N = 292$ ; EukDis = Euklidische Distanz innerhalb eines Testlets; QuaDiff =

Summe quadrierter Differenzen innerhalb eines Testlets;  $r_s$  4 Items = Spearman

Rangkorrelation innerhalb eines Testlets (d. h. über 4 Items) gebildet;  $r_s$  16 Items = Spearman

Rangkorrelation über 16 Items gebildet. Für den EAP-Test werden Itemstatistiken vor und nach der Itemselektion berichtet.

In Tabelle 5.1 werden zudem für die unterschiedlichen Maße der Akkuratheit im DEF-Test Itemstatistiken berichtet. Die beiden Distanz-Maße weisen zufriedenstellende Trennschärfen auf, wobei die Schätzungen etwas besser für die Euklidische Distanz ausfallen, als die für die Summe quadrierter Differenzen. Die Schätzungen für die Trennschärfen der korrelativen Maße fallen insgesamt etwas schlechter aus. Die Variante, bei der die Rangkorrelation über 16 Items gebildet wird, weist deutlich bessere Trennschärfen auf, als die Variante, über 4



Items. Für beide korrelativen Maße konnten durch Versuche, eine Itemselektion durchzuführen, keine bedeutsamen Verbesserungen der Messgenauigkeit erzielt werden.

### 5.3.1 Deskriptive Statistiken und Reliabilitätsschätzungen

In Tabelle 5.2 werden deskriptive Statistiken und Reliabilitätsschätzungen für den EAP-Test sowie für die unterschiedlichen Maße des DEF-Tests dargestellt. Die Schätzungen wurden jeweils getrennt für die Partnerin bzw. den Partner 1 und 2 vorgenommen. Auch in dieser Studie wurden die Reliabilität des EAP-Tests über 6 Testlets als Analyseeinheit berichtet, anstelle über die einzelnen Items, um das Problem der Abhängigkeit der Items zu berücksichtigen. Mit Schätzungen von  $\alpha = .78$  für Partnerin bzw. Partner 1 und  $\alpha = .74$  für Partnerin bzw. Partner 2 sind die Reliabilitätsschätzungen zufriedenstellend und liegen in einer Höhe, die mit den Ergebnissen aus Studie 1 vergleichbar sind.

Für den DEF-Test fallen die Reliabilitätsschätzungen für die Distanzmaße und die korrelativen Maße sehr unterschiedlich aus. Für die euklidischen Distanzen als Maß der Akkuratheit liegt Cronbachs  $\alpha$  bei .71 und .68 und für die Summe quadrierter Differenzen bei .73 und .69. Die Schätzungen für Cronbachs  $\alpha$  für die beiden Distanzmaße liegen somit in einer vergleichbaren Höhe und können im Hinblick auf die Ziele dieser Studie als zufriedenstellend betrachtet werden.

Im Gegensatz dazu fallen die Reliabilitätsschätzungen für beide Varianten der Rangkorrelation als Maß der Akkuratheit gering aus. Für die Variante der Rangkorrelation über 4 Items liegen die Schätzungen für Cronbachs  $\alpha$  bei .12 und .28. Für die Variante mit Rangkorrelationen über 16 Items liegt die Schätzung von Cronbachs  $\alpha$  etwas höher, mit jeweils  $\alpha = .33$  für beide Sets von Partnerinnen und Partner. Beide Varianten der Rangkorrelation weisen dennoch Schätzungen für Cronbachs  $\alpha$  auf, die als nicht akzeptabel zu bewerten sind.

Aufgrund dieser Ergebnisse wurden im weiteren Verlauf dieser Studie die beiden Distanzmaße den korrelativen Maßen vorgezogen. Aufgrund der Schätzungen für die Verteilungsparameter wurde zwischen den beiden Distanzmaßen die Euklidische Distanz als Maß der Akkuratheit gegenüber der Summe Quadrierter Differenzen bevorzugt. Da korrelative Maße einen anderen Aspekt der Übereinstimmung betonen, werden die Ergebnisse für das bessere der beiden Maße (d. h. die Rangkorrelation über 16 Items) im Folgenden ebenfalls berichtet, wobei dies allein einer Gegenüberstellung dient und Schlussfolgerungen des korrelativen Maßes nur vorsichtig getroffen werden.

Tabelle 5.2

*Deskriptive Statistiken und Reliabilitätsschätzungen für den EAP-Test sowie für die Genauigkeits-Maße des DEF-Tests jeweils für Partner/in 1 und Partner/in 2*

Maß	Analyseeinheit	<i>M</i>	<i>SD</i>	Schiefe	Kurtosis	$\alpha$
EAP-Test						
P1	6 Testlets	0.64	0.12	-0.07	0.48	.78
P2	6 Testlets	0.63	0.11	0.29	0.49	.74
DEF-Test						
EukDis P1	20 Testlets	5.97	1.10	0.67	0.49	.71
EukDis P2	20 Testlets	5.89	1.06	0.43	-0.27	.68
QuaDiff P1	20 Testlets	43.72	15.94	1.09	1.19	.73
QuaDiff P2	20 Testlets	42.94	15.02	0.71	0.12	.69
$r_s$ 4 Items P1	20 Testlets	0.40	0.15	-0.07	0.11	.12
$r_s$ 4 Items P2	20 Testlets	0.42	0.15	-0.25	0.15	.28
$r_s$ 16 Items P1	5 Indikatoren	0.45	0.14	-0.38	0.61	.33

Maß	Analyseeinheit	<i>M</i>	<i>SD</i>	Schiefe	Kurtosis	$\alpha$
$r_s$ 16 Items P2	5 Indikatoren	0.47	0.14	-0.71	0.86	.33

*Anmerkungen.* P1 = Partner/in 1; P2 = Partner/in 2;  $n_{P1}$  = 146;  $n_{P2}$  = 146; EukDis = Euklidische Distanz innerhalb eines Testlets; QuaDiff = Summe quadrierter Differenzen innerhalb eines Testlets;  $r_s$  4 Items = Spearman Rangkorrelation innerhalb eines Testlets (d. h. über 4 Items) gebildet;  $r_s$  16 Items = Spearman Rangkorrelation über 16 Items gebildet. Die Schätzungen für Cronbachs  $\alpha$  basieren auf Testlets bzw. den angegebenen Indikatoren.

### 5.3.2 Strukturprüfung des DEF-Tests: dyadische CFA

Für den DEF-Test wurde die angenommene eindimensionale Struktur mit einer dyadischen CFA geprüft, wie sie von Olsen und Kenny (2006) für ununterscheidbare Dyaden vorgeschlagen wurde. Für das euklidische Distanzmaß wurden die 20 Itemgruppen zufällig zu 5 Parcels zugeordnet. Die gleiche Zuordnung von Items zu Parcels wurde für Partner 1 und 2 vorgenommen, so dass die Parcels für Partner 1 und 2 die gleichen Items enthalten. Als Schätzer für die CFA wurde ML verwendet. Um die Voraussetzung multivariater Normalverteilung zu prüfen, wurden die unvariaten Verteilungen der Parcels über eine Inspektion von Histogrammen und Verteilungsparameter der Indikatoren betrachtet. Univariat konnten keine größeren Abweichungen von der Normalverteilung festgestellt werden. In Abbildung 5.3 werden die Ergebnisse der Parameterschätzungen für die dyadische CFA zur Testung der eindimensionalen Struktur des DEF-Tests unter Verwendung der euklidischen Distanz präsentiert.

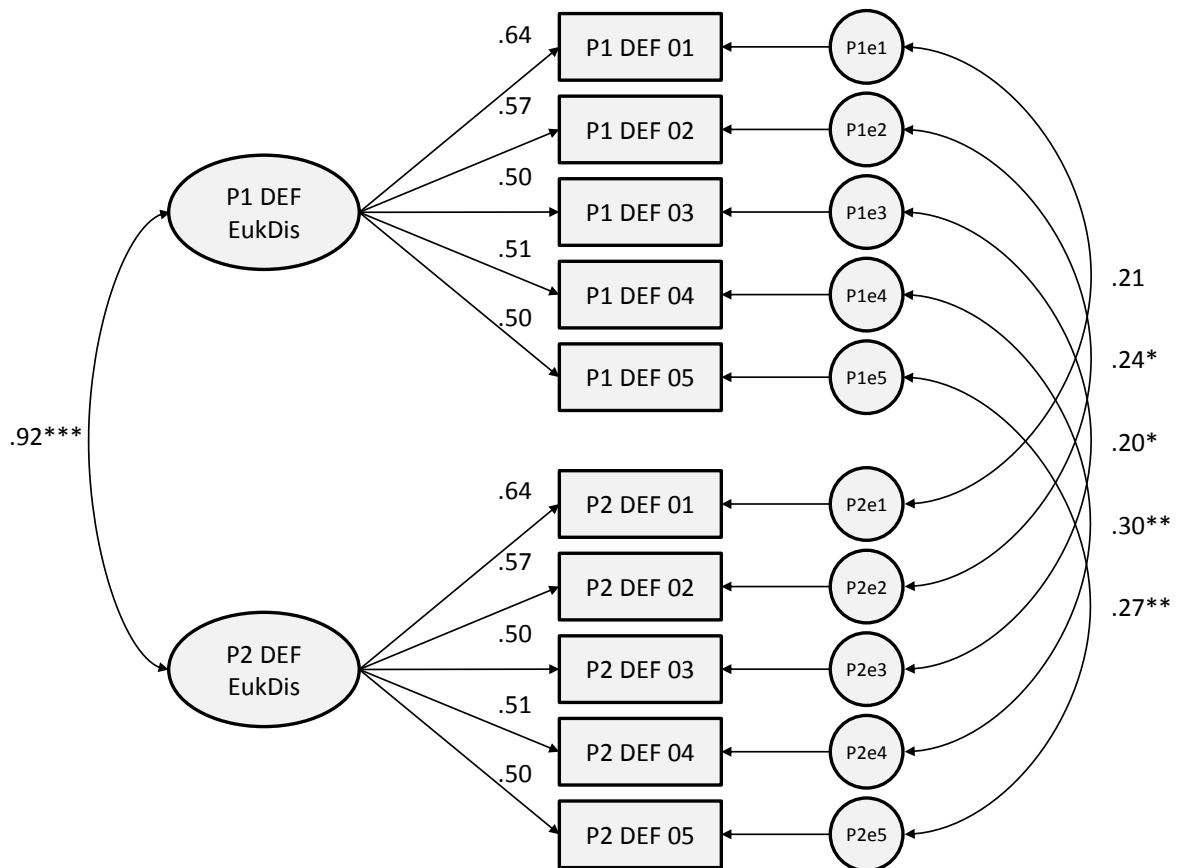


Abbildung 5.3. Dyadische CFA für den DEF-Test mit der Euklidischen Distanz als Maß der Übereinstimmung. Sämtliche Faktorladungen sind auf dem .001-Niveau signifikant. \* =  $p < .05$ , \*\* =  $p < .01$ , \*\*\* =  $p < .001$ .

In Tabelle 5.3 können die Fitstatistiken für das Modell entnommen werden. Für den DEF-Test mit der euklidischen Distanz als Maß der Akkuratheit sprechen die Ergebnisse dafür, dass das angenommene eindimensionale Modell sehr gut auf die Daten passt. Die anhand des I-SAT und I-NUL Modells adjustierten globalen Fit-Indizes weisen mit  $CFI = .98$  und  $RMSEA = .06$  auf eine sehr gute Passung des eindimensionalen Modells hin. Zudem sind sämtliche Faktorladungen auf dem .001-Niveau signifikant, so dass der lokale Fit des Modells ebenfalls gegeben ist.

Tabelle 5.3

*Fit-Statistiken für die dyadische CFA des DEF-Tests mit Euklidischer Distanz und Rangkorrelation über 16 Items als Maße der Akkuratheit*

	$\chi^2$	df	$\chi^2'$	df'	<i>p</i>	CFI	RMSEA
<b>EukDiss</b>							
Modell	47.55	44	20.59	14	.11	.98	.06
I-SAT	26.96	30					
I-NULL	368.11	55					
<b><i>r<sub>s</sub></i> 16 Items</b>							
Modell	53.77	44	14.35	14	.42	1.00	.01
I-SAT	39.42	30					
I-NULL	204.94	55					

*Anmerkungen.* Der Modelltest und die Fit-Indizes basieren auf dem anhand des I-SAT Modells korrigierten  $\chi^2$ -Wert ( $\chi^2'$ ) und Freiheitsgraden (df') sowie dem I-NULL Modell.

Zum Vergleich werden in Abbildung 5.4 die Parameterschätzungen für die dyadische CFA zur Prüfung der eindimensionalen Struktur des DEF-Tests unter Verwendung der Rangkorrelation als Maß der Akkuratheit berichtet. Die adjustierten globalen Fit-Indizes weisen auf einen sehr guten Modell-Fit hin. Allerdings sind eine Reihe der Faktorladungen nicht signifikant von Null verschieden.

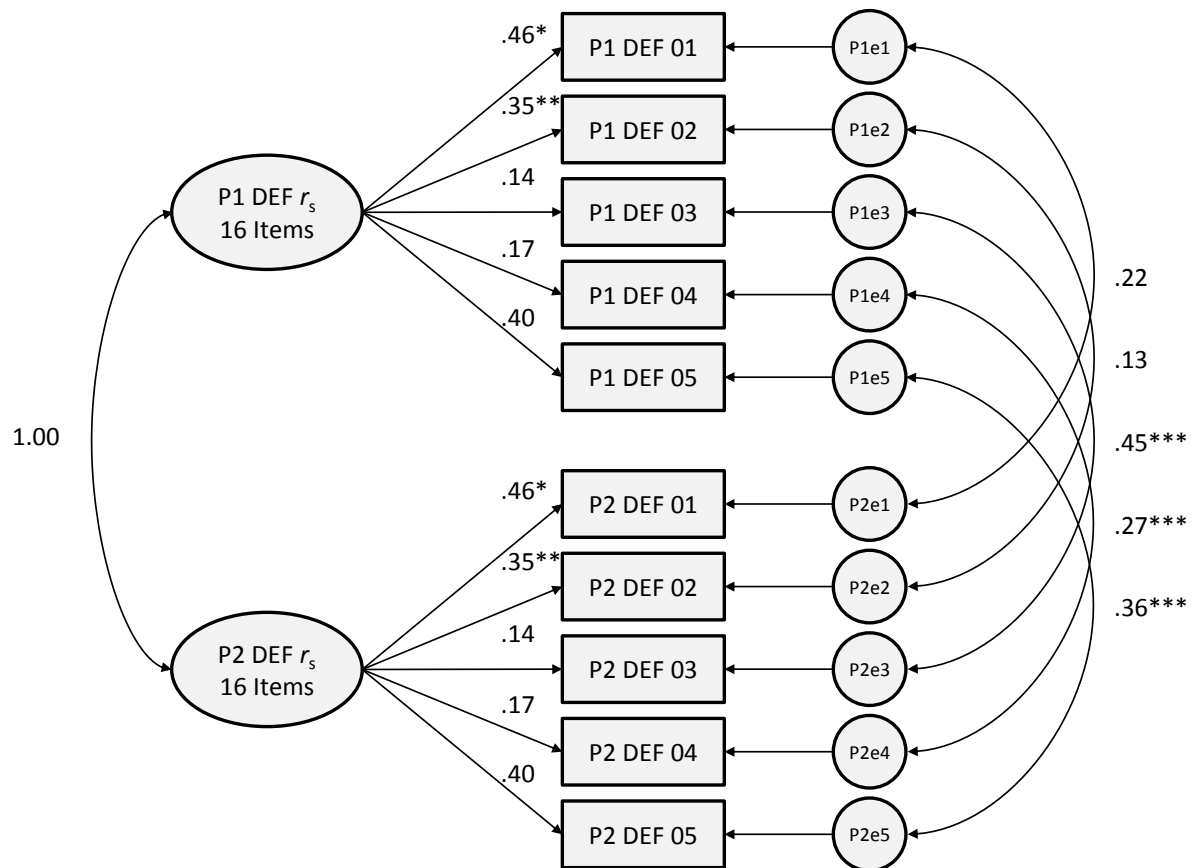


Abbildung 5.4. Dyadische CFA für den DEF-Test mit Rangkorrelationen über 16 Items als Maß der Übereinstimmung. \* =  $p < .05$ , \*\* =  $p < .01$ , \*\*\* =  $p < .001$ .

### 5.3.4 Actor-Partner Interdependence Modell

In Abbildung 5.3 wird das Actor-Partner Interdependence Modell mit Distanzen als Maß der Akkuratheit im DEF-Test gezeigt. Der Zusammenhang zwischen Testleistung im EAP-Test und DEF-Test wird im Actor-Effekt widerspiegelt. Erwartungskonform besteht ein negativer Zusammenhang zwischen der Leistung im EAP-Test und der Leistung im DEF-Test, gemessen über die euklidische Distanz. Allerdings ist der standardisierte Regressionskoeffizient mit  $-.11$  deutlich geringer als erwartet. Die Partner-Effekte sind in

diesem Modell nicht signifikant von Null verschieden. Die Abhängigkeit der Leistungen der Partnerinnen und Partner spiegelt sich in der deutlichen Korrelation zwischen den Leistungen im EAP-Test (.37) sowie der hohen Korrelation der Residuen wider (.70).

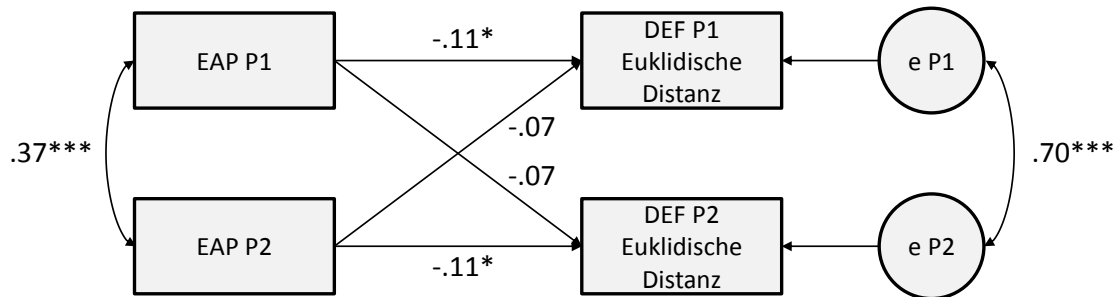


Abbildung 5.5. APIM für die Regression der Leistung im DEF-Test (euklidische Distanz) auf die Leistung im EAP-Test. P1 = Partner/in 1; P2 = Partner/in 2. \* =  $p < .05$ , \*\*\* =  $p < .001$ . Modell-Fit:  $\chi^2_{ML} = 5.45$ ,  $df = 6$ ,  $p = .49$ , CFI = 1.00, RMSEA = .00. Fit-Statistiken sind für das I-SAT Modell identisch.

In Abbildung 5.6 wird das APIM mit der Rangkorrelation als Maß für den DEF-Test zur Möglichkeit des Vergleichs gegenübergestellt. Interessanterweise zeigt sich hier trotz der geringen Reliabilität ein höherer erwartungskonformer Zusammenhang zwischen EAP-Test und DEF-Test. In diesem Modell ist auch der Partnereffekt mit .22 signifikant von Null verschieden. Zwar erscheinen die Zusammenhänge des DEF-Tests mit dem EAP-Test hier besser, aufgrund der ungenügenden Reliabilität sowie der Auffälligkeiten bei der CFA sollte dieses Ergebnis vorsichtig betrachtet werden.

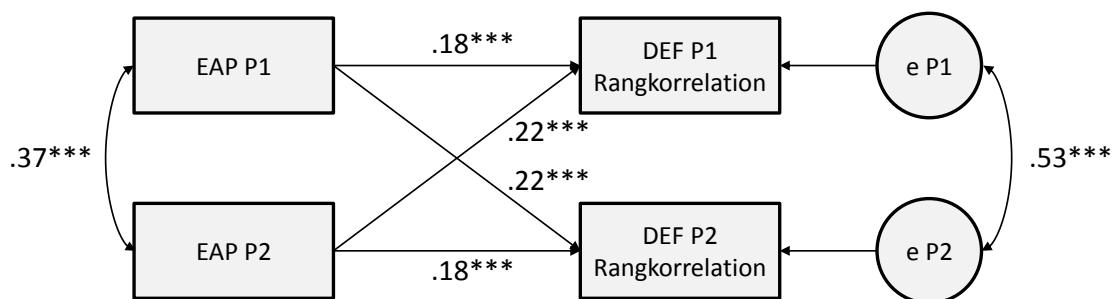


Abbildung 5.6. APIM für die Regression der Leistung im DEF-Test (Rangkorrelation über 16 Items) auf die Leistung im EAP-Test. P1 = Partner/in 1; P2 = Partner/in 2.  $*** = p < .001$ . Modell-Fit:  $\chi^2_{ML} = 4.37$ ,  $df = 6$ ,  $p = .63$ , CFI = 1.00, RMSEA = .00. Fit-Statistiken sind für das I-SAT Modell identisch.

### 5.3.5 Zusammenhänge mit verbaler fluider Intelligenz und den Big-Five Persönlichkeitsfaktoren

Um die Beziehungen des DEF-Tests mit fluider verbaler Intelligenz im Sinne konvergenter Validitätsevidenz zu untersuchen, wurden die Korrelationen der Maße des DEF-Tests mit dem BEFKI Schlussfolgerndes Denken – verbal und dem 24-AM betrachtet. In Tabelle 5.4 werden deskriptive Statistiken und Schätzungen der internen Konsistenz für den BEFKI und die Big-Five Persönlichkeitsfaktoren getrennt für Partnerin bzw. Partner 1 und 2 berichtet. Hinsichtlich der internen Konsistenz weist die Skala für Gf verbal des BEFKI mit  $\alpha = .68$  und  $\alpha = .71$  zufriedenstellende Schätzungen auf. Für den 24-AM variieren die Schätzungen für Cronbachs  $\alpha$ . Offenheit für Erfahrung und Verträglichkeit weisen sehr geringe interne



Konsistenzen auf. Für die restlichen Faktoren können die Schätzungen für die Reliabilität als zufriedenstellend bis gut bezeichnet werden.

Tabelle 5.4

*Deskriptive Statistiken und Reliabilitätsschätzungen für die Instrumente für Gf-verbal und die Big-Five Persönlichkeitsfaktoren*

Test	Anzahl Items	P1			P2		
		<i>M</i>	<i>SD</i>	$\alpha$	<i>M</i>	<i>SD</i>	$\alpha$
BEFKI Gf verbal	16	9.60	2.72	.68	9.73	2.77	.71
O	5	5.17	0.80	.45	4.96	0.79	.31
C	5	5.22	1.14	.76	5.24	1.08	.72
E	6	4.88	1.18	.83	4.91	1.07	.80
A	4	4.83	0.76	.43	4.71	0.95	.64
N	4	3.49	1.24	.73	3.42	1.27	.72

*Anmerkungen.* P1 = Partner/in 1; P2 = Partner/in 2; Gf = fluide Intelligenz; O = Offenheit für Erfahrung; C = Gewissenhaftigkeit; E = Extraversion; A = Verträglichkeit; N = Neurotizismus.

In Tabelle 5.5 werden die Zusammenhänge zwischen DEF-Test und EAP-Test mit Gf-verbal und den Big-Five Persönlichkeitsfaktoren berichtet. Entgegen den Erwartungen für ein Fähigkeitsmaß, weist der DEF-Test unter Verwendung des Distanzmaßes keine Beziehung zu schlussfolgerndem Denken – verbal auf. Der DEF-Test mit Rangkorrelation als Maß der Akkuratheit weist hingegen erwartungskonforme positive Zusammenhänge mit dem

Intelligenzmaß auf. Der EAP-Test weist, wie erwartet, positive substantielle Korrelationen mit Schlussfolgerndem Denken – verbal auf. Darüber hinaus ist der EAP-Test, wie in Studie 1, mit Offenheit für Erfahrung und mit Verträglichkeit korreliert und unabhängig von den übrigen Persönlichkeitsfaktoren. Für den DEF-Test hingegen können die erwarteten Beziehungen mit Offenheit für Erfahrung nicht beobachtet werden. Hier zeigen sich unerwartete Beziehungen mit Extraversion sowie mit Verträglichkeit.

Tabelle 5.5

*Korrelative Zusammenhänge zwischen DEF-Test und EAP-Test mit Gf-verbal und den Big-Five Persönlichkeitsfaktoren getrennt für Partnerin bzw. Partner 1 und 2*

	Gf verbal	O	C	E	A	N
P1: DEF-Test EukDis	-.11	.01	.12	-.11	-.25*	.12
P2: DEF-Test EukDis	-.18	-.02	-.16	-.29*	-.21	.11
P1: DEF-Test $r_s$ 16 Items	.29*	.09	-.13	-.10	.32**	-.09
P2: DEF-Test $r_s$ 16 Items	.24*	.15	.08	.25*	.22	-.22
P1: EAP	.59**	.12	-.11	-.14	.10	-.11
P2: EAP	.57**	.24*	-.14	.12	.24*	-.03

*Anmerkungen.* Zusammenhänge mit Gf – verbal basieren auf einer Stichprobengröße von  $n = 71$  und Zusammenhänge mit den Big-Five Persönlichkeitsfaktoren auf  $n = 75$ . P1 = Partner/in 1; P2 = Partner/in 2; EukDis = Euklidische Distanz innerhalb eines Testlets;  $r_s$  16 Items = Spearman Rangkorrelation über 16 Items gebildet. Gf = fluide Intelligenz; O = Offenheit für Erfahrung; C = Gewissenhaftigkeit; E = Extraversion; A = Verträglichkeit; N = Neurotizismus.\* =  $p < .05$ , \*\* =  $p < .01$ .

## 5.4 Diskussion

Die dritte Studie wurde mit dem Ziel durchgeführt, weitere Validitätsbelege für den EAP-Test zu finden, indem der theoretisch erwartete Zusammenhang des EAP-Tests mit einem neu konstruierten Instrument zur Erfassung von Empathic Accuracy untersucht wurde. Insgesamt muss konstatiert werden, dass die Ergebnisse dieser Studie eher enttäuschend ausfallen. Die Zusammenhänge des EAP-Tests mit den Maßen des DEF-Tests sprechen nur mit Einschränkungen für die Validität des EAP-Tests. Zwar werden im Rahmen der APIM sowohl für das Distanzmaß als auch für das korrelative Maß des DEF-Tests Beziehungen mit dem EAP-Test beobachtet, die die erwartete Richtung aufweisen. Die Höhe der Zusammenhänge ist allerdings deutlich geringer als erwartet, wenn man die hohe konzeptuelle Überlappung zwischen EU und Empathic Accuracy bedenkt. Die substantiellen positiven Beziehungen zwischen EAP-Test und schlussfolgerndem Denken – verbal hingegen replizieren einen Validitätsbeleg aus Studie 1, der mit den Maßen fluider Intelligenz des I-S-T 2000 R gefunden wurden. Zudem weist der EAP-Test erneut die erwarteten Beziehungen zu den Big-Five Persönlichkeitsfaktoren auf, was als Replikation des diskriminanten Validitätsbelegs aus Studie 1 betrachtet werden kann.

Weitere Schlussfolgerungen über die Validität des EAP-Tests können anhand der vorliegenden Ergebnisse nur schwer gezogen werden. Die Gründe für die unerwartet geringen Beziehungen zwischen EAP-Test und DEF-Test könnten in der möglicherweise mangelnden Validität des EAP-Tests, des DEF-Tests oder in beiden Tests gleichzeitig liegen. Die Betrachtung der korrelativen Beziehungen des DEF-Tests mit Maßen für Intelligenz und den Big-Five Persönlichkeitsfaktoren, werfen weitere Zweifel hinsichtlich der Validität des DEF-Tests auf. Im Sinne des positive manifold Phänomens sollte die Leistung im DEF-Test positiv mit einem etablierten Maß für Intelligenz korreliert sein. Das Distanzmaß für den DEF-Test weist allerdings in diesen Daten keinen Zusammenhang mit schlussfolgerndem Denken –

verbal auf. Interessanterweise weist das korrelative Maß des DEF-Tests eine positive moderate Beziehung zum Schlussfolgernden Denken – verbal auf, wie sie für einen Leistungstest erwartet werden kann. Dieses Ergebnis ist erstaunlich, zumal das korrelative Maß des DEF-Tests im Gegensatz zum Distanzmaß ungenügende psychometrische Eigenschaften aufweist. Aus diesen Ergebnissen wird ersichtlich, dass der Bestimmung der Akkuratheit im DEF-Test weitere Beachtung geschenkt werden sollte.

In zukünftigen Studien könnte eine Verbesserung der Bestimmung der Empathic Accuracy angestrebt werden. Dabei könnten die von Cronbach (1955) vorgeschlagenen Ansätze zur Bestimmung sinnvoller Maße für empathische Fähigkeiten Hinweise für Verbesserungen der Messung liefern. Cronbach (1955) hat eine Zerlegung der Komponenten von Übereinstimmungsmaßen vorgeschlagen. Eine dieser Komponenten, die als Differential Accuracy bezeichnet wurde, wurde von Cronbach (1955) als sinnvolles Maß für die Bestimmung empathischer Akkuratheit vorgeschlagen. Differential Accuracy ist die Fähigkeit einer Person, Unterschiede zwischen den Zielpersonen bezüglich eines Items zu identifizieren (siehe Cronbach, 1955). Die Bestimmung dieses Maßes erfordert allerdings, dass ein Beobachter verschiedene Zielpersonen einschätzt (siehe Cronbach, 1955), so dass die Bestimmung des Maßes im Rahmen der verfügbaren Daten nicht möglich war.

Zudem muss bedacht werden, dass es sich beim DEF-Test um einen Versuch der Abwandlung des Empathic Accuracy Ansatzes handelt. Ein Grund für mögliche Schwächen des DEF-Tests könnte darin liegen, dass die vorgegebenen Situationen von den Zielpersonen noch nie erlebt wurden oder schwer vorgestellt werden können. Wie es generell beim Target-Scoring der Fall ist, ist auch beim DEF-Test zu kritisieren, dass das Kriterium die Selbstauskunft einer Zielperson ist. Es ist daher zu erwarten, dass die Güte der Messung auch davon abhängig ist, inwieweit die jeweiligen Zielpersonen ihre eigenen emotionalen

Reaktionen für die vorgegebenen Situationen vorhersagen können und/oder Auskunft darüber geben wollen.

Trotz der eher enttäuschenden Ergebnisse hinsichtlich der erwarteten Beziehungen zwischen EAP-Test und DEF-Test, sollte der Ansatz, die Validität eines EU-Tests mithilfe von Ansätzen aus der Empathie- oder Emotionsforschung zur prüfen, nicht verworfen werden. Die hohe konzeptuelle Überschneidung und, mit Einschränkungen, auch die Ergebnisse dieser Studie sprechen dafür, dass solche Maße interessante Variablen zur Validierung für Tests von EU liefern können.

## 6. Diskussion

### 6.1 Zusammenfassende Bewertung des EAP-Tests

In dieser Arbeit wurde, basierend auf dem von Schulze und Roberts (Schulze et al., 2009) vorgeschlagenen EAP, ein neues Verfahren zur Erfassung von EU entwickelt, dessen psychometrische Qualität und Validität in drei Studien untersucht wurde. Insgesamt sind die Ergebnisse für den neu entwickelten EAP-Test vielversprechend. Sowohl hinsichtlich der Reliabilität als auch der Validität des EAP-Tests konnte eine Reihe wichtiger Belege erbracht werden. Bezüglich der Reliabilität wies der EAP-Test in allen drei Studien Schätzungen für Cronbachs  $\alpha$  auf, die für einen ersten Testentwurf in einem akzeptablen Bereich lagen ( $\alpha = .70$  bis  $.78$ ), deren Verbesserung dennoch das Ziel zukünftiger Optimierungen des Tests darstellen sollten. Im Zusammenhang mit den Reliabilitätsschätzungen, wurde der Testlet-Problematik besonderes Augenmerk geschenkt. Die potentielle Gefahr der Verletzung der Unabhängigkeit der Items sowie die damit zusammenhängende Gefahr der Überschätzung der Reliabilität (Wainer et al., 2007), wurde hier durch zwei Maßnahmen berücksichtigt: Zum einen wurde in Studie 1 das Ausmaß der Abhängigkeit der Items im EAP-Test untersucht, indem das von Wang et al. (2002) vorgeschlagene Testlet-Modell an die Daten angepasst wurde. Es konnte gezeigt werden, dass die aufgrund der Struktur des Tests vermutete fehlende Unabhängigkeit der Items tatsächlich beobachtet werden kann und damit ein Problem darstellt. Sowohl für Items, die mit einer Emotionskontingenz gepaart sind, als auch für Itemgruppen, die in ein Szenario genestet sind, wurden substantielle Testlet-Effekte identifiziert. Zum anderen wurde in allen drei Studien bei der Ermittlung von Cronbachs  $\alpha$  auf 6 Testlets, die den 6 Szenarien entsprechen, als Analyseeinheit zurückgegriffen, was in Anbetracht der vorhandenen Testlet-Effekte konservativere Schätzungen der Reliabilität erlaubte. Die Verwendung von Testdesigns, bei denen mehrere Items auf den gleichen Gegenstand bezogen werden, ist allerdings kein Alleinstellungsmerkmal des EAP-Tests. Insbesondere für die, in der Domäne

EI anzutreffenden SJTs, in denen mehrere Abfragen mit einer Situationsbeschreibung gepaart sind, ist von der gleichen Problematik auszugehen. Zu diesen SJTs zählen beispielsweise einige Untertests des MSCEIT (Mayer et al., 2002), der STEM (MacCann & Roberts, 2008) sowie der TEMINT (Amelang & Steinmayr, 2006). In zukünftigen Studien sollte auch für diese Tests bzw. Untertests das Problem möglicher Verletzungen der Unabhängigkeitsannahme der Items untersucht werden und bei Schlussfolgerungen über die Messgenauigkeit der Tests berücksichtigt werden, was bisher nur in wenigen Ausnahmen der Fall war (Maul, 2012).

Bezüglich der Validität des EAP-Tests konnten im Rahmen der drei Studien mehrere Belege gefunden werden, auf deren Basis eine zusammenfassende Schlussfolgerung über die Validität des EAP-Tests, im Sinne des Validität-Konzeptes von Messick (1995), ermöglicht werden sollte. In Studie 1 konnte mittels einer CFA die angenommene eindimensionale Struktur des EAP-Tests erfolgreich belegt werden. Die Passung des Modells konnte dabei für zwei Varianten von Parcels (d. h. sowohl über 14 Itemgruppen als auch über 6 Szenarien als Indikatoren) stabil als sehr gut bezeichnet werden. Dieses Ergebnis stellt einen grundlegenden Validitätsbeleg dar, da mit dem EAP-Test ein einzelnes Konstrukt gemessen werden soll.

Um sich anschließend einer Antwort auf die Frage anzunähern, ob es sich bei dem einen Konstrukt tatsächlich um EU handelt, wurden konvergente und diskriminante Validitätsbelege hinzugezogen. In Studie 1 konnten die aus dem nomologischen Netzwerk erwarteten Beziehungen des EAP-Tests sowohl zu klassischer Intelligenz als auch den Big-Five Persönlichkeitsfaktoren weitestgehend bestätigt werden. Der EAP-Test wies in Studie 1 die erwarteten positiven Beziehungen zu klassischer Intelligenz auf. In Studie 3 konnten die positiven Beziehungen mit einem weiteren Instrument für klassische Intelligenz repliziert werden. Bezüglich der theoretisch angenommenen besonderen Nähe des EAP-Tests zu Maßen für Gf im Vergleich zu Gc konnten in Studie 1 erste Hinweise gefunden werden, die im

Einklang mit Konzeptualisierungen von EU als Schlussfolgerndes Denken über Emotionen stehen (z. B. Mayer et al., 2001). Diese Ergebnisse sprechen allerdings gegen die genannte, häufige Auffassung von EU als Form von Gc (Davies et al., 1998; MacCann, 2010; Steinmayr et al., 2011; Zeidner, et al., 2008). In der deutschsprachigen Fassung des MSCEIT wurde der Branch EU sogar als Emotionswissen benannt (Steinmayr et al., 2011). Anhand der Höhe der Faktorkorrelationen zwischen EAP-Test und den Maßen klassischer Intelligenz kann im Sinne eines diskriminanten Validitätsbelegs zudem davon ausgegangen werden, dass die mit dem EAP-Test gemessene Fähigkeit von klassischer Intelligenz unterscheidbar ist.

Der EAP-Test wies darüber hinaus ein Muster von Faktorkorrelationen zu den Big-Five Persönlichkeitsfaktoren auf, das den Erwartungen für einen Test zur Messung einer Form von Intelligenz entspricht (siehe Ackerman & Heggestad, 1997). Insgesamt war der EAP-Test relativ unabhängig von den Big-Five Persönlichkeitsfaktoren, mit Ausnahme einer moderat positiven Beziehung zu Offenheit für Erfahrung. Auch dieses erwartungskonforme Ergebnis konnte in Studie 3 mit einem weiteren Instrument für die Big-Five weitestgehend repliziert werden. Einschränkend muss für Studie 3 erwähnt werden, dass aus Gründen der Durchführbarkeit nur eine Kurzsкала der Big-Five zum Einsatz kam, deren Reliabilitätsschätzungen für die Persönlichkeitsfaktoren Offenheit für Erfahrung und Verträglichkeit recht gering waren. Im Zusammenhang mit den Big-Five Persönlichkeitsfaktoren konnte darüber hinaus ein geringer positiver Zusammenhang zwischen dem EAP-Test und dem Persönlichkeitsfaktor Verträglichkeit beobachtet werden. Zwar wurde ein solcher Zusammenhang auch für andere fähigkeitsbasierte EI-Tests identifiziert (siehe Roberts et al., 2008); aus Sichtweise klassischer Intelligenz-Diagnostik stellt dies nichtsdestotrotz ein ungewöhnliches Ergebnis dar (Ackerman & Heggestad, 1997), das näherer Untersuchung bedarf.



Um einen weiteren konvergenten Validitäts-Beleg für den EAP-Test finden zu können, wurde in Studie 1 der TBEU als theoriebasiertes Maß für EU entwickelt. Der EAP-Test wies erwartungskonform zum TBEU die höchste positive Beziehung auf. Da es sich um ein neu entwickeltes Testverfahren handelt, wurden auch für den TBEU Belege bezüglich seiner Qualität gesammelt. Mit einer Schätzung von Cronbachs  $\alpha$  von .76 erwies sich der TBEU als hinreichend reliabel für diese Untersuchung. Zudem konnte die angenommene eindimensionale Struktur des TBEU über drei Parcel-Varianten hinweg belegt werden. Basierend auf einer Teil-Stichprobe von Studie 1, konnte auch ein erster Hinweis für ein Muster an Beziehungen zu Intelligenz und Persönlichkeitsfaktoren gefunden werden, das zum größten Teil den aus der Theorie abgeleiteten Erwartungen entsprach. Insgesamt muss aber trotz der positiven Befunde für den TBEU konstatiert werden, dass dem zugrundeliegenden Messansatz das Problem der Abhängigkeit von der Richtigkeit der dahinterliegenden Theorie anhaftet, wie es in dieser Arbeit bereits für den STEU (MacCann & Roberts, 2008) kritisiert wurde. Aus dieser Perspektive ist der EAP-Test dem TBEU klar zu bevorzugen. Dennoch wurde mit dem TBEU ein weiteres Verfahren zur Verfügung gestellt, dass gerade im Hinblick auf Validierungsstudien durch die Verwendung eines alternativen Konstruktionsansatzes hilfreich sein kann.

Mit Studie 2 konnte experimentelle Validitätsevidenz für den EAP-Test geliefert werden. Die Annahme, dass die Testpersonen ihre Vorhersagen tatsächlich auf Grundlage von Informationen aus der Aneignungsphase treffen und für die korrekte Beantwortung nicht ebenso gut auf allgemeines Wissen zurückgreifen können, konnte in dieser Studie bestätigt werden. Dieser Beleg ist wichtig, da auf Grundlage dieser Annahme das vergleichsweise aufwändige Testdesign mit Aneignungs- und Anwendungsphase sowie das Scoring begründet werden. Darüber hinaus konnte gezeigt werden, dass Items besonders dann schwer zu lösen sind, wenn die Zielpersonen Emotionskontingenzen aufweisen, die zuvor als ungewöhnlich eingeschätzt wurden. Dieses Ergebnis liefert einen wertvollen Hinweis auf ein Merkmal der

Items, mit dem die Itemschwierigkeit im Rahmen der Testkonstruktion mit dem EAP gezielt beeinflusst werden kann.

In Studie 3 wurde die Beziehung des EAP-Tests mit einem Messansatz aus der Empathie-Forschung untersucht. Der in diesem Zusammenhang neu entwickelte DEF-Test wurde hierbei als eine Form von Kriteriums-Messung betrachtet, mit der die Fähigkeit, Wissen über Emotionskontingenzen von realen Personen zu erwerben, erfasst werden sollte. Da es sich um ein neues Instrument handelte, wurde der DEF-Test ebenfalls hinsichtlich seiner Reliabilität und Validität überprüft. Zwar wurden Zusammenhänge zwischen EAP-Test und den Maßen der Akkuratheit des DEF-Tests gefunden, die den erwarteten Richtungen entsprachen, allerdings fielen diese deutlich geringer aus, als theoretisch anzunehmen war. Eine Limitation der Aussagekraft dieser Studie liegt in der Qualität des DEF-Tests begründet. Aufgrund der inakzeptablen Reliabilitätsschätzungen für den DEF-Test sowie der teilweise mangelnden konvergenten Validitätsbelege mit klassischer Intelligenz, können Schlussfolgerung über den EAP-Test anhand des DEF-Tests nur schwierig gezogen werden. Trotz dieser Limitation kann zunächst ausschließlich festgehalten werden, dass in der 3. Studie nur teilweise Evidenz für den EAP-Test gefunden werden konnte.

Über die hier berichteten empirischen Belege hinaus, gibt es auch inhaltliche Gründe, die für die Validität des EAP-Tests sprechen. An bisherigen Tests für EI wurde kritisiert, dass die Auswahl von Items nicht näher begründet wurde und das Scoring zweifelhaft ist (z. B., Schulze et al., 2007). Mit dem Ziel einer Verbesserung der Testentwicklung haben Schulze et al. (2007) die Verwendung des Evidence Centered Design (ECD) als Rahmenmodell vorgeschlagen, anhand dessen jeder Schritt der Testentwicklung und die damit verbundenen Entscheidungen explizit gemacht werden könnten. Wie im Abschnitt zur Messung von EI erwähnt, wurde die Konstruktion der SJTs in der Domäne der EI selten näher beschrieben und begründet. Bei der Konstruktion des EAP-Tests bot das EAP ein Prinzip, das ein

systematischeres und begründetes Vorgehen beim Erstellen von Items erlaubte und zudem ein eindeutiges Scoring-Kriterium bei der Bewertung der Testantworten bot. Das EAP bietet ein Prinzip, mit dem das Erkennen von Regelmäßigkeiten, das einen zentralen Aspekt der Theorie von EU darstellt, messbar gemacht werden kann. Bei der Entwicklung des EAP-Tests anhand des EAP wurden weitere Entscheidungen getroffen, die die Validität des Instruments sicherstellen sollten. So wurde versucht, über das konkrete Design der Items potentielle Konfundierungen mit nicht-intendierten Konstrukten zu kontrollieren. Zum einen wurde versucht, den potentiellen Einfluss von Merkfähigkeit weitestgehend zu reduzieren. Die Faktorkorrelation zwischen EAP-Test und Merkfähigkeit weist darauf hin, dass dies gelungen ist. Zum anderen wurde der Einfluss von Perception kontrolliert, indem die Items so gestaltet wurden, dass das Wahrnehmen von Emotionen für die korrekte Beantwortung der Items keine Rolle spielt.

Mit dem EAP muss nicht, wie bei bisherigen Tests, auf Scoring-Methoden zurückgegriffen werden (d. h. Konsensus-, Experten- oder Target-Scoring), deren Adäquatheit in dieser Arbeit anhand einer Befunde angezweifelt wurde (siehe auch Roberts et al., 2001). Anders als beim theoriebasierten Scoring, hängt die Angemessenheit des Scorings im EAP auch nicht von der Richtigkeit einer bestimmten Theorie ab. Die einzige Annahme, die im EAP getroffen werden muss ist, dass Emotionen Regelmäßigkeiten unterliegen, die individuell verschieden sein können und dass Personen dies nutzen können, um Emotionen vorherzusagen. Aktuelle Theorien zum Emotionslöseprozess, wie die Appraisal-Theorien, erlauben diese individuell verschiedenen Regelmäßigkeiten. Auch Theorien der Persönlichkeit, in denen eine Stabilität des Verhaltens angenommen wird, erscheinen vereinbar mit dem EAP (z. B. Friesen & Kammrath, 2011).

Insgesamt betrachtet liefern die ersten beiden Studien, und, mit Einschränkungen, auch die dritte Studie Belege für die psychometrische Qualität und Validität des EAP-Tests.

Zudem sprechen auch inhaltliche Gesichtspunkte für die Angemessenheit des hier verwendeten Testdesigns und Scorings.

## **6.2 Zukünftige Studien zum EAP-Test**

In zukünftigen Studien sollte eine Verbesserung des EAP-Tests auf Grundlage der bisherigen Ergebnisse anzielt werden sowie weitere Belege für dessen Validität gesammelt werden. Zwar sind die ersten Ergebnisse für den EAP-Test zufriedenstellend, allerdings ist das Instrument hinsichtlich einiger Aspekte auch verbesserungswürdig. Zum einen sollte die Reliabilität des Tests verbessert werden, indem neue Items generiert werden, die die ausgewählten Items ergänzen können. Zudem enthält der EAP-Test bisher deutlich mehr leichte als schwere Items. Damit mit dem EAP-Test auch im hohen Fähigkeitsbereich differenziert werden kann, sollte die Erstellung schwerer Items angestrebt werden. Bei der Konstruktion neuer Items sollten Merkmale der Aufgaben, von denen vermutet wird oder bekannt ist, dass sie die Schwierigkeit beeinflussen können, systematischer variiert werden, um Einflüsse auf die Schwierigkeit testen zu können. In diesem Zusammenhang könnten auch Itemkomponenten-Modelle im Rahmen der IRT hilfreich sein (siehe Rost, 2004), um den Einfluss von Itemmerkmalen auf die Schwierigkeit modellieren zu können. Einen Hinweis auf ein Merkmal, das die Schwierigkeit von Items im EAP erhöht, liefert Studie 2. Dabei ist im Moment noch nicht klar, wann Emotionskontingenzen als gewöhnlich oder ungewöhnlich wahrgenommen werden, was sich vermutlich auch in der eher geringen Übereinstimmung der Beurteiler in Studie 2 gezeigt hat. Vielleicht könnten hier Appraisal-Theorien, wie vorgeschlagen wurde, wiederum Hinweise dafür liefern, was gewöhnliche oder ungewöhnliche Reaktionen sein könnten (Orchard et al., 2009). Im EAP-Test konnten bisher nur Items generiert werden, bei denen die emotionale Reaktion eine Wahrscheinlichkeit von 0 oder 1 besitzt. Eine Möglichkeit für Weiterentwicklungen des EAP-Tests besteht darin, differenziertere Wahrscheinlichkeiten zu modellieren, was die Schwierigkeit insgesamt

erhöhen könnte und gleichzeitig die Ratewahrscheinlichkeit, die bisher bei 50 % liegt, senken kann. Hierzu müssten die Zielpersonen allerdings in mehr als einer Situation beobachtet werden, was hinsichtlich der damit verbundenen Testverlängerung problematisch ist. Des Weiteren könnte versucht werden, die Schwierigkeiten der Items durch einen größeren Transfer zwischen Aneignungs- und Anwendungsphase zu erhöhen (Orchard et al., 2009). Für die Eindeutigkeit des Scoring muss hierbei allerdings begründet werden können, hinsichtlich welcher Merkmale die Situation in der Aneignungs- und Anwendungsphase vergleichbar sind.

Bezüglich der Validitätsevidenz konnte in den vorliegenden Studien nur eine eng begrenzte Auswahl von Maßen für klassische Intelligenz eingesetzt werden. Die Schlussfolgerung, dass der Einfluss von Merkfähigkeit eher gering ist, beruht bisher nur auf den verbalen Aufgaben des I-S-T 2000 R. In weiteren Untersuchungen wäre es sinnvoll, Merkfähigkeitsaufgaben zu verwenden, die zum einen messgenauer sind und zum anderen verschiedene Aspekte der Merkfähigkeit erfassen. Beispielsweise wäre eine Abgrenzung zu figuraler Merkfähigkeit ebenfalls sinnvoll, da das Testdesign so gestaltet ist, dass Zielpersonen u. a. durch Form und Farbe unterschieden werden können. Diese Abgrenzung von Merkfähigkeit bleibt auch für die Validierung zukünftiger Revisionen des EAP-Tests wichtig. Denkbar wäre nämlich, dass sich die Beziehungen zu Merkfähigkeit verändern, wenn z. B. im Rahmen einer Verbesserung des EAP-Tests neue Items entwickelt werden. Beispielsweise könnte eine starke Erhöhung der Informationen, die kurzfristig behalten werden muss, dazu führen, dass der Einfluss von Merkfähigkeit deutlich stärker wird, als es in dieser Studie beobachtet wurde.

In zukünftigen Studien sollte zudem der Zusammenhang zwischen dem EAP-Test und Maßen für klassische Intelligenz näher untersucht werden. Sowohl die hier gefundenen Zusammenhänge als auch theoretische Annahmen über EU legen nahe, dass es sich bei EU um eine Form der fluiden Intelligenz handelt. Andererseits sprechen die hier genannten

Ergebnisse für andere Tests zur Erfassung von EU auch für eine Nähe zu Gc (siehe MacCann et al., 2011). Es wäre denkbar, dass es neben dem Schlussfolgernden Denken über Emotionen auch eine Form von allgemeinem Wissen über Emotionen gibt. Diese Form von EU könnte Wissen über Emotionsbegriffe beinhalten, wie es im Rahmen des Four-Branch Modells als ein Aspekt von EU vorgeschlagen wurde (Mayer & Salovey, 1997; Mayer et al., 2000; siehe auch Abschnitt Understanding in Kapitel 2.1.2).

Zudem sollten in zukünftigen Studien weitere kriterienbezogene Belege für die Validität des EAP-Tests erbracht werden. Bei der Wahl geeigneter Kriterien sollte vorab begründet werden können, warum anzunehmen ist, dass das Kriterium durch die Fähigkeit EU vorhergesagt werden kann. Für Maße klassischer Intelligenz stellt akademischer Erfolg, beispielsweise über Schulnoten operationalisiert, ein sinnvolles Kriterium dar (z. B. Jäger, Holling, Preckel, Schulze, Vock, Süß & Beauducel, 2006; Liepmann et al., 2006). Akademischen Erfolg auch als Kriterium für den EAP-Test heranzuziehen, wäre nur dann sinnvoll, wenn begründet werden kann, inwiefern EU für akademischen Erfolg entscheidend ist. Ein reines Explorieren, ob der EAP-Test mit Schulnoten oder wünschenswerten Eigenschaften, wie Gesundheitsverhalten, zusammenhängt, erscheint wenig zielführend. Bezüglich kriterienorientierter Validitätsbelege muss darüber hinaus auch betrachtet werden, inwiefern der EAP-Test über etablierte Maße der Intelligenz hinaus Vorhersagekraft besitzt.

### **6.3 Weitere Anwendungen für das EAP**

Das EAP ist von Schulze et al. (2009) insbesondere für die Entwicklung von Tests für EU vorgeschlagen worden. Das EAP könnte allerdings auch auf andere Fähigkeiten angewendet werden, die ebenfalls das Verstehen von Regelhaftigkeiten beinhalten. Beispielsweise könnte der Bereich der Diagnostik sozialer Intelligenz, der vor ähnlichen Problemen wie die Diagnostik von EI steht, vom EAP profitieren. Insbesondere für die Entwicklung von Tests für Social Understanding (z. B. Weis & Süß, 2005), könnte das EAP hilfreich sein. Auch

wenn Social Understanding ohnehin eine hohe konzeptuelle Überschneidung mit EU aufweist, könnte das EAP auf andere Inhalte erweitert werden, wie zum Beispiel die Motivationen von Zielpersonen. Eine weitere Möglichkeit besteht in der Anwendung im Bereich der Fähigkeit, die Persönlichkeitseigenschaften anderer Personen akkurat einschätzen zu können, der sogenannten Judgmental Accuracy (siehe Colvin & Bundick, 2001). Über das EAP könnten Zielpersonen mit bestimmten Persönlichkeitseigenschaften modelliert werden, deren Verhalten die Testperson aufgrund des vorherigen Verhaltens vorhersagen soll. In diesem Zusammenhang wäre auch eine Verwendung des EAP im Rahmen der neu vorgeschlagenen *Personalen Intelligenz* (Mayer, 2007-2008) denkbar, die als die Fähigkeit zum Schlussfolgern über Persönlichkeitseigenschaften definiert wurde.

#### **6.4 Fortschritt im Bereich EI**

Für den Fortschritt im Bereich EI ist die Verfügbarkeit von Messverfahren und sinnvollen Scoring-Ansätzen notwendig. Nur unter diesen Umständen können Theorien der EI getestet und ggf. weiterentwickelt werden. Das Four-Branch Modell hat einen wichtigen Beitrag für die Erforschung von EI geleistet. Allerdings weisen einige Befunde bereits jetzt darauf hin, wie bspw. die schwache Evidenz für den Faktor Facilitation (siehe Fan et al., 2010), dass das Modell zukünftig überarbeitet werden muss. Vor dem Hintergrund dieser notwendigen Überarbeitung ist es wünschenswert, dass EI nicht mehr fast ausschließlich mit einem einzigen Messansatz, d. h. dem MSCEIT, erfasst wird.

EI steht darüber hinaus bisher eher isoliert zu nah verwandten Forschungsbereichen, wie der Emotions- und Empathieforschung. Nur wenige Studien (z. B. Roberts et al., 2006; MacCann et al., 2011) haben bislang Maße für EI anhand von etablierten Maßen aus der Emotionsforschung validiert. Für die Messung und Entwicklung der Theorie zu EU könnten vor allem Ansätze aus der Forschung zur fähigkeitsbasierten Empathie hilfreich sein.

Daneben sollte EI zukünftig auch stärker mit Theorien der klassischen Intelligenz in Verbindung gebracht werden. Dazu sollte versucht werden, das Four-Branch Modell in etablierte Intelligenz-Modelle zu integrieren. Bislang sind Vorhersagen über die Beziehungen der Branches zu klassischer Intelligenz teilweise schwierig, da einige der Branches, wie im Falle von EU, widersprüchlich konzeptualisiert wurden. Eine Möglichkeit, EI in klassische Intelligenztheorien verorten zu können, könnten Facettenmodelle der Intelligenz bieten, die um eine Inhaltskomponente der Emotionen ergänzt werden könnte. Ähnliches wurde bereits im Bereich sozialer Intelligenz versucht (Conzelmann et al., 2013).

Die Entwicklung von Mess- und Scoring-Ansätzen für EI stellt auch weiterhin ein wichtiges Ziel in dieser Domäne dar, ohne das Fortschritt nicht möglich sein wird. Mit dieser Arbeit konnte Evidenz für das von Schulze und Roberts (Schulze et al., 2009) vorgeschlagene EAP gesammelt werden und gleichzeitig ein Instrument zur Verfügung gestellt werden, mit dem EU zukünftig erfasst werden kann.



## Zusammenfassung

Emotional Understanding (EU) stellt eine von vier Fähigkeiten dar, die im Rahmen des sogenannten Four-Branch Modells der Emotionalen Intelligenz (EI) vorgeschlagen wurden (z. B. Mayer et al., 2008). Obwohl EI seit ihrer Einführung in den 1990ern hohe Aufmerksamkeit erhalten hat, ist der Forschungsstand in dieser Domäne nicht so weit vorangeschritten, wie angesichts der Bemühungen erwartet werden könnte. Stattdessen steht die Forschung zur EI teilweise noch vor einigen sehr grundlegenden Fragen, die die Existenz des Konstruktes und die Möglichkeit zur Operationalisierung betreffen. Zwei der schwerwiegendsten Probleme im Bereich der EI, die aktuell den Forschungsprozess verlangsamten, sind in a) konzeptuellen Unklarheiten sowie b) im Fehlen eines validen Scoring-Kriteriums zur Bewertung von richtigen und falschen Testantworten zu finden. Auf konzeptueller Ebene besteht das Problem darin, dass es neben dem Four-Branch Modell der EI, das auf Fähigkeiten begrenzt ist, eine Reihe weiterer Modelle gibt, in denen unter dem Begriff EI verschiedene Konzepte aus der Persönlichkeitsdomäne mit emotionsbezogenen Fähigkeiten vermischt werden (z. B. Schulze et al., 2007). Zudem können in den unterschiedlichsten psychologischen Domänen Konzepte identifiziert werden, wie z. B. Empathie, die hohe Ähnlichkeit mit EI bzw. EU aufweisen, aber selten im Rahmen der EI betrachtet wurden.

Hinsichtlich der Messung von EI kann das Phänomen beobachtet werden, dass die Zahl der Selbstberichtsverfahren die Anzahl verfügbarer Leistungstests deutlich übersteigt. Die geringe Anzahl von Leistungstests sowie die Tatsache, dass Erkenntnisse über EI fast ausschließlich auf einem einzelnen Messansatz, dem MSCEIT bzw. MEIS (Mayer et al., 2002), beruhen, stellen keinen wünschenswerten Zustand dar. Zudem sind die wenigen Leistungstests zur Messung von EI bzw. EU mit Problemen behaftet. Aufgrund des Mangels

an einem veridikalen Scoring-Kriteriums, musste bislang auf Scoring-Methoden zurückgegriffen werden, die angezweifelt wurden (z. B. Roberts et al., 2001).

Mit dem Ziel der Verbesserung der Messung von EU, entwickelten Ralf Schulze und Richard D. Roberts das Empathic Agent Paradigma (EAP) (Schulze et al., 2009). Im Rahmen des EAP sind Tests durch ein Zwei-Phasen-Design charakterisiert: In der ersten Phase (der Aneignungsphase) besteht die Aufgabe der Testperson darin, die typischen emotionalen Reaktionen einer oder mehrerer Zielpersonen auf ein bestimmtes Ereignis zu beobachten. In der zweiten Phase (der Aneignungsphase) muss die Testperson das Wissen über die typischen emotionalen Reaktionen der Zielperson aus der Aneignungsphase in einer neuen, vergleichbaren Situation korrekt anwenden. In dieser Arbeit wurde unter Verwendung des EAP ein neuer, computerbasierter Test maximalen Verhaltens für EU entwickelt: der EAP-Test. Durch bestimmte Maßnahmen bei der Entwicklung des Testdesigns des EAP-Tests wurde insbesondere versucht, den Einfluss der Emotionswahrnehmungsfähigkeit auf die Testleistung zu kontrollieren.

In drei Studien wurde die psychometrische Qualität sowie Belege für die Validität des EAP-Tests untersucht. Bei der Untersuchung der Reliabilität wurde das Problem potentieller Testlet-Effekt berücksichtigt, die aufgrund der Struktur des Tests angenommen wurden (Wang et al., 2002). Tatsächlich konnten Belege für Testlet-Effekte nach Anpassung des Testlet-Modells von Wang et al. (2002) erbracht werden. Für den EAP-Test wurden in den drei Studien akzeptable Reliabilitätsschätzungen unter Berücksichtigung der Testlet-Effekte gefunden ( $\alpha = .70$  bis  $.78$ ). In Studie 1 ( $N = 210$ ) wurde der EAP-Test zusammen mit Maßen für klassische Intelligenz, für die Big-Five Persönlichkeitsfaktoren und mit einem neu entwickelten theoriebasierten Test für EU präsentiert. Auf Ebene latenter Variablen wurden Zusammenhänge beobachtet, die für einen Test der EU erwartet werden können.

In Studie 2 wurde experimentelle Validitätsevidenz für den EAP-Test erbracht. In dieser Untersuchung wurden die Teilnehmerinnen und Teilnehmer ( $N = 158$ ) zufällig zwei Test-Bedingungen zugewiesen: Der EAP-Test wurde entweder ohne die Aneignungsphase (Experimentalgruppe) oder mit der Aneignungsphase (Kontrollgruppe) präsentiert. Wie erwartet erzielte die Experimentalgruppe geringere Testwerte als die Kontrollgruppe, der im Gegensatz zur Experimentalgruppe vorab Informationen über die Zielpersonen zur Verfügung standen. Dieser Effekt war zudem erwartungskonform höher für Items, in denen die Zielpersonen ungewöhnliche emotionale Reaktionen aufwiesen. Die Ergebnisse unterstützen die Annahme, dass Testpersonen bei der Vorhersage von emotionalen Reaktionen der Zielpersonen auf erworbenes Wissen aus der Aneignungsphase zurückgreifen und nicht allein auf Grundlage von allgemeinem Wissen über Emotionen eine Einschätzung treffen.

In der dritten Studie wurde der EAP-Test zusammen mit einem neu entwickelten Test für Empathic Accuracy (siehe z. B. Ickes, 2001) in einer dyadischen Stichprobe von 146 Paaren untersucht. Der EAP-Test und der Test für Empathic Accuracy wiesen eine positive Beziehung zueinander auf, die allerdings deutlich geringer ausfiel als erwartet.

Insgesamt sprechen die Ergebnisse dafür, dass das EAP ein nützliches Werkzeug für die Konstruktion und das Scoring von Tests der EU ist. Mit dem EAP-Test konnte zudem ein neues, vielversprechendes Instrument zur Erfassung von EU vorgeschlagen werden.

## Referenzen

- Ackerman, P. L. & Heggestad, E. D. (1997). Intelligence, personality, and interests: evidence for overlapping traits. *Psychological Bulletin*, *121*, 219—245. doi: 10.1037/0033-2909.121.2.219
- Allen, V. D., Weissman, A., Hellwig, S., MacCann, C. & Roberts, R. D. (2014). Development of the Situational Test of Emotional Understanding – Brief (STEU-B) using item response theory. *Personality and Individual Differences*, *65*, 3—7. doi: 10.1016/j.paid.2014.01.051
- Amelang, M. & Steinmayr, R. (2006). Is there a validity increment for tests of emotional intelligence in explaining the variance of performance criteria? *Intelligence*, *34*, 459—468. doi: 10.1016/j.intell.2006.03.003
- Austin, E. J. (2010). Measurement of ability emotional intelligence: results for two new tests. *British Journal of Psychology*, *101*, 563—578. doi: 10.1348/000712609X474370
- Bänziger, T., Grandjean, D. & Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT). *Emotion*, *9*, 691—704. doi: 10.1037/a0017088
- Bänziger, T., Mortillaro, M. & Scherer, K. R. (2012). Introducing the Geneva Multimodal Expression Corpus for experimental research on emotion perception. *Emotion*, *12*, 1161—1179. doi: 10.1037/a0025827
- Banse, R. & Scherer K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*, 614—636. doi: 10.1037//0022-3514.70.3.614
- Barrett, L. F., Mesquita, B. & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, *20*, 286—290. doi: 10.1177/0963721411422522

- Barchard, K. A., Hensley, S. & Anderson, E. (2013). When proportion consensus scoring works. *Personality and Individual Differences*, 55, 14—18. doi: 10.1016/j.paid.2013.01.017
- Barchard, K. A. & Russell, J. A. (2006). Bias in consensus scoring, with examples from ability emotional intelligence tests. *Psicothema*, 18, 49—54. Retrieved from <http://www.psicothema.com/pdf/3275.pdf>
- Bar-On, R. (2006). The Bar-On Model of emotional-social intelligence (ESI). *Psicothema*, 18, 13—25. Retrieved from [www.psicothema.com/pdf/3271.pdf](http://www.psicothema.com/pdf/3271.pdf)
- Bar-On, R., Brown, J. M., Kirkcaldy, B. D. & Thomé, E. P. (2000). Emotional expression and implications for occupational stress; an application of the Emotional Quotient Inventory (EQ-i). *Personality and Individual Differences*, 28, 1107—1118. doi: 10.1016/S0191-8869(99)00160-9
- Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R. & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, 29, 407—418. doi: 10.1023/A:1023035012436
- Baron-Cohen, S. & Wheelwright, S. (2004). The Empathy Quotient: An investigation of adults with Asperger Syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, 34, 163—175. doi: 10.1023/B:JADD.0000022607.19833.00
- Batson, C. D. (2009). Two forms of perspective taking: Imagining how another feels and imagining how you would feel. In K. D. Markman, W. M. P. Klein & J. A. Suhr (Eds.), *Handbook of Imagination and Mental Simulation* (S. 267—279). New York: Psychology Press.

- Bernhardt, B. C. & Singer, T. (2012). The neural basis of empathy. *Annual Review of Neuroscience*, *35*, 1—23. doi: 10.1146/annurev-neuro-062111-150536
- Blanchette, I. & Richards, A. (2010). The influence of affect on higher level cognition: A review of research on interpretation, judgement, decision making and reasoning. *Cognition & Emotion*, *24*, 561—595. doi: 10.1080/02699930903132496
- Blickle, G., Momm, T. S., Kramer, J., Mierke, J., Liu, Y. & Ferris, G. R. (2009). Construct and criterion-related validation of a measure of emotional reasoning skills: a two-study investigation. *International Journal of Selection and Assessment*, *17*, 101—118. doi: 10.1111/j.1468-2389.2009.00455.x
- Blickle, G., Momm, T., Liu, Y., Witzki, A. & Steinmayr, R. (2011). Construct validation of the Test of Emotional Intelligence (TEMINT). *European Journal of Psychological Assessment*, *27*, 282—289. doi: 10.1027/1015-5759/a000075
- Bonanno, G. A. & Burton, C. L. (2013). Regulatory flexibility: An individual differences perspective on coping and emotion regulation. *Perspectives on Psychological Science*, *8*, 591—612. doi: 10.1177/1745691613504116
- Brackett, M. A. & Mayer, J. D. (2003). Convergent, discriminant, and incremental validity of competing measures of emotional intelligence. *Personality and Social Psychology Bulletin*, *29*, 1147—1158. doi: 10.1177/0146167203254596
- Brody, N. (2004). What cognitive intelligence is and what emotional intelligence is not. *Psychological Inquiry*, *15*(3), 234—238.
- Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81—105. doi: 10.1037/h0046016

- Cannon, W. B. (1927). The James-Lange Theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology*, *39*, 106—124. doi: 10.2307/1415404
- Carpenter, P. A., Just, M. A. & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*, 404—431. doi: 10.1037/0033-295X.97.3.404
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: a critical experiment. *Journal of Educational Psychology*, *54*, 1—22. doi: 10.1037/h0046743
- Cheung, M. W.-L. & Chang, W. (2005). Meta-analytic structural equation modeling: a two-stage approach. *Psychological Methods*, *10*, 40—64. doi: 10.1037/1082-989X.10.1.40
- Chien Farh, C. I. C., Seo, M. G. & Tesluk, P. E. (2012). Emotional intelligence, teamwork effectiveness, and job performance: the moderating role of job context. *Journal of Applied Psychology*, *97*, 890—900. doi: 10.1037/a0027377
- Christian, M. S., Edwards, B. D. & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*, 83—117. doi: 10.1111/j.1744-6570.2009.01163.x
- Ciarrochi, J. V., Chan, A. Y. C. & Caputi, P. (2000). A critical evaluation of the emotional intelligence construct. *Personality and Individual Differences*, *28*, 539—561. doi: 10.1016/S0191-8869(99)00119-1
- Colvin, C. R. & Bundick, M. J. (2001). In search of the good judge of personality: some methodological and theoretical concerns. In J. A. Hall & F. Bernieri (Eds.),

- Interpersonal sensitivity: theory and measurement* (S. 47—65). Mahwah, NJ: Lawrence Erlbaum Associates.
- Conzelmann, K., Weis, S. & Süß, H.-M. (2013). New findings about social intelligence: Development and application of the Magdeburg Test of Social Intelligence (MTSI). *Journal of Individual Differences, 34*, 119—137. doi: 10.1027/1614-0001/a000106
- Cook, W. L. & Kenny, D. A. (2005). The Actor-Partner Interdependence Model: A model of bidirectional effects in developmental studies. *International Journal of Behavioral Development, 29*, 101—109. doi: 10.1080/01650250444000405
- Costa, A., Faria, L. & Takšić, V. (2011). Vocabulary of Emotions Test (VET): a cross-cultural validation in Portuguese and Croatian contexts. *Psihologijske Teme, 20*(3), 413—424.
- Costa, P. T., Jr. & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13*, 653—665. doi: 10.1016/0191-8869(92)90236-I
- Côté, S., DeCelles, K. A., McCarthy, J. M., Van Kleef, G. A. & Hideg, I. (2011). The Jekyll and Hyde of emotional intelligence: Emotion-regulation knowledge facilitates both prosocial and interpersonally deviant behavior. *Psychological Science, 22*, 1073—1080. doi: 10.1177/0956797611416251
- Cronbach, L. J. (1955). Processes affecting scores on “understanding of others” and “assumed similarity”. *Psychological Bulletin, 52*, 177—193. doi: 10.1037/h0044919
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5. Aufl.). New York: Harper Collins Publishers.
- Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281—302. doi: 10.1037/h0040957



- Dael, N., Mortillaro, M. & Scherer, K. R. (2012). Emotion expression in body action and posture. *Emotion, 12*, 1085—1101. doi: 10.1037/a0025737
- Davidson, J. E. & Downing, C. L. (2000). Contemporary models of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (S. 34—49). Cambridge: Cambridge University Press.
- Davies, M., Stankov, L. & Roberts, R. D. (1998). Emotional intelligence: In search of an elusive construct. *Journal of Personality and Social Psychology, 75*, 989—1015. doi: 10.1037/0022-3514.75.4.989
- Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*, 113—126. doi: 10.1037/0022-3514.44.1.113
- Davis, M. H., Mitchell, K. V., Hall, J. A., Lothert, J., Snapp, T. & Meyer, M. (1999). Empathy, expectations, and situational preferences: Personality influences on the decision to participate in volunteer helping behaviors. *Journal of Personality, 67*, 469—503. doi: 10.1111/1467-6494.00062
- Davis, S. K. & Humphrey, N. (2012). Emotional intelligence predicts adolescent mental health beyond personality and cognitive ability. *Personality and Individual Differences, 52*, 144—149. doi: 10.1016/j.paid.2011.09.016
- Dawda, D. & Hart, S. D. (2000). Assessing emotional intelligence: Reliability and validity of the Bar-On Emotional Quotient Inventory (EQ-i) in university students. *Personality and Individual Differences, 28*, 797—812. doi: 10.1016/S0191-8869(99)00139-7

- Day, A. L. & Carroll, S. A. (2008). Faking Emotional Intelligence (EI): comparing response distortion on ability and trait-based EI measures. *Journal of Organizational Behavior*, 29, 761—784. doi: 10.1002/job.485
- Decety, J. (2010). To what extent is the experience of empathy mediated by shared neural circuits? *Emotion Review*, 2, 204—207. doi: 10.1177/1754073910361981
- Digman, J. M. & Inouye, J. (1986). Further specification of the five robust factors of personality. *Journal of Personality and Social Psychology*, 50, 116—123. doi: 10.1037/0022-3514.50.1.116
- Dunn, E. W., Brackett, M. A., Ashton-James, C., Schneiderman, E. & Salovey, P. (2007). On emotionally intelligent time travel: Individual differences in affective forecasting ability. *Personality and Social Psychology Bulletin*, 33, 85—93. doi: 10.1177/0146167206294201
- Dunn, J., Brown, J., Slomkowski, C., Tesla, C. & Youngblade, L. (1991). Young children's understanding of people's feelings and beliefs: individual differences and their antecedents. *Child Development*, 62, 1352—1366. doi: 10.1111/j.1467-8624.1991.tb01610.x
- Dymond, R. F. (1949). A scale for the measurement of empathic ability. *Journal of Consulting Psychology*, 13, 127—133. doi: 10.1037/h0061728
- Dziobek, I., Rogers, K., Fleck, S., Bahnemann, M., Heekeren, H. R., Wolf, O. T. & Convit, A. (2008). Dissociation of cognitive and emotional empathy in adults with Asperger syndrome using the Multifaceted Empathy Test (MET). *Journal of Autism and Developmental Disorders*, 38, 464—473. doi: 10.1007/s10803-007-0486-x

- Ekman, P. & Oster, H. (1979). Facial expressions of emotion. *Annual Review of Psychology*, *30*, 527—544. doi: 10.1146/annurev.ps.30.020179.002523
- Empathie. (n.d.). In *Duden*. Retrieved from <http://www.duden.de/rechtschreibung/Empathie>
- Eysenck, H. J. (1992). Four ways five factors are not basic. *Personality and Individual Differences*, *13*, 667—673. doi: 10.1016/0191-8869(92)90237-J
- Fan, Y., Duncan, N. W., de Greck, M. & Northoff, G. (2011). Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neuroscience & Biobehavioral Reviews*, *35*, 903—911. doi: 10.1016/j.neubiorev.2010.10.009
- Fan, H., Jackson, T., Yang, X., Tang, W. & Zhang, J. (2010). The factor structure of the Mayer-Salovey-Caruso Emotional Intelligence Test V 2.0 (MSCEIT): a meta-analytic structural equation modeling approach. *Personality and Individual Differences*, *48*, 781—785. doi: 10.1016/j.paid.2010.02.004
- Farrelly, D. & Austin, E. J. (2007). Ability EI as an intelligence? Associations of the MSCEIT with performance on emotion processing and social tasks and with cognitive ability. *Emotion & Cognition*, *21*, 1043—1063. doi: 10.1080/02699930601069404
- Ferguson, F. J. & Austin, E. J. (2010). Associations of trait and ability emotional intelligence with performance on theory of mind tasks in an adult sample. *Personality and Individual Differences*, *49*, 414—418. doi: 10.1016/j.paid.2010.04.009
- Ford, M. E. & Tisak, M. S. (1983). A further search for social intelligence. *Journal of Educational Psychology*, *75*, 196—206. doi: 10.1037/0022-0663.75.2.196
- Freudenthaler, H. H. & Neubauer, A. C. (2005). Emotional intelligence: The convergent and discriminant validities of intra- and interpersonal emotional abilities. *Personality and Individual Differences*, *39*, 569—579. doi: 10.1016/j.paid.2005.02.004

- Freund, P. A. & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, *138*, 296—321. doi: 10.1037/a0026556
- Friesen, C. A. & Kammrath, L. K. (2011). What it pays to know about a close other: The value of if-then personality knowledge in close relationships. *Psychological Science*, *22*, 567—571. doi: 10.1177/0956797611405676
- Gardner, H. (1991). *Abschied vom IQ: die Rahmen-Theorie der vielfachen Intelligenz*. Stuttgart: Klett-Cotta.
- Gardner, H. & Hatch, T. (1989). Educational implications of the theory of multiple intelligence. *Educational Researcher*, *18*, 4—10. doi: 10.3102/0013189X018008004
- Goldberg, L. R. (1990). An alternative “description of personality”: the Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*, 1216—1229. doi: 10.1037/0022-3514.59.6.1216
- Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ*. New York, NY: Bantam Books.
- Gosling, S. D., Rentfrow, P. J. & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504—528. doi: 10.1016/S0092-6566(03)00046-1
- Gross, J. J. (1998). The emerging field of emotion regulation: an integrative review. *Review of General Psychology*, *2*, 271—299. doi: 10.1037/1089-2680.2.3.271
- Gross, J. J. (2008). Emotion regulation. In M. Lewis, J. M. Haviland-Jones & L. F. Barrett (Eds.), *Handbook of emotion* (3. Aufl., S. 497—512). New York: The Guilford Press.

- Gross, J. J. & John, O. P. (2003). Individual differences in two emotion regulation processes: implication for affect, relationships, and well-being. *Journal of Personality and Social Psychology, 85*, 348—362. doi: 10.1037/0022-3514.85.2.348
- Grynberg, D., Luminet, O., Corneille, O., Grèzes, J. & Berthoz, S. (2010). Alexithymia in the interpersonal domain: a general deficit of empathy? *Personality and Individual Differences, 49*, 845—850. doi: 10.1016/j.paid.2010.07.013
- Guilford, J. P. (1959). Three faces of intellect. *American Psychologist, 14*, 469—479. doi: 10.1037/h0046827
- Guilford, J. P. (1966). Intelligence: 1965 model. *American Psychologist, 21*, 20—26. doi: 10.1037/h0023296
- Guilford, J. P. (1988). Some changes in the Structure-of-Intellect Model. *Educational and Psychological Measurement, 48*, 1—4. doi: 10.1177/001316448804800102
- Guttman, L. (1959). A structural theory for intergroup beliefs and action. *American Sociological Review, 24*, 318—328. doi: 10.2307/2089380
- Guttman, R. & Greenbaum, C. W. (1998). Facet theory: its development and current status. *European Psychologist, 3*, 13—36. doi: 10.1027/1016-9040.3.1.13
- Hakstian, A. R. & Cattell, R. B. (1978). Higher-stratum ability structures on a basis of twenty primary abilities. *Journal of Educational Psychology, 70*, 657—669. doi: 10.1037/0022-0663.70.5.657
- Harris, P. L. (2008). Children's understanding of emotion. In M. Lewis, J. M. Haviland-Jones & L. F. Barrett (Eds.), *Handbook of emotion* (3. Aufl., S. 320—331). New York: The Guilford Press.

- Hart, B. & Spearman, C. (1912). General ability, its existence and nature. *British Journal of Psychology*, 5, 51—84. doi: 10.1111/j.2044-8295.1912.tb00055.x
- Herzberg, P. Y. & Brähler, E. (2006). Assessing the Big-Five personality domains via short forms: a cautionary note and a proposal. *European Journal of Psychological Assessment*, 22, 139—148. doi: 10.1027/1015-5759.22.3.139
- Hoerger, M., Chapman, B. P., Epstein, R. M. & Duberstein, P. R. (2012). Emotional intelligence: A theoretical framework for individual differences in affective forecasting. *Emotion*, 12, 716—725. doi: 10.1037/a0026724
- Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, 57, 253—270. doi: 10.1037/h0023816
- Horn, J. L. & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica*, 26, 107—129. doi: 10.1016/0001-6918(67)90011-X
- Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1—55. doi: 10.1080/10705519909540118
- Huber, J., Ariely, D. & Fischer, G. (2002). Expressing preferences in a principal-agent task: a comparison of choice, rating, and matching. *Organizational Behavior and Human Decision Processes*, 87, 66—90. doi: 10.1006/obhd.2001.2955
- Hughes, C. & Dunn, J. (1998). Understanding mind and emotion: longitudinal associations with mental-state talk between young friends. *Developmental Psychology*, 34, 1026—1037. doi: 10.1037/0012-1649.34.5.1026

- Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology*, *60*, 653—670. doi: 10.1146/annurev.psych.60.110707.163604
- Ickes, W. (2001). Measuring empathic accuracy. In J. A. Hall & F. Bernieri (Eds.), *Interpersonal sensitivity: theory and measurement* (S. 219—241). Mahwah, NJ: Lawrence Erlbaum Associates.
- Inquisit (Version 3) [Computer software]. Seattle, WA: Millisecond Software.
- Izard, C., Fine, S., Schultz, D., Mostow, A., Ackerman, B. & Youngstrom, E. (2001). Emotion knowledge as a predictor of social behavior and academic competence in children at risk. *Psychological Science*, *12*, 18—23. doi: 10.1111/1467-9280.00304
- Izard, C., Trentacosta, C., King, K., Morgan, J. & Diaz, M. (2007). Emotions, emotionality, and intelligence in the development of adaptive behavior. In G. Matthews, M. Zeidner & R. D. Roberts (Eds.), *The science of emotional intelligence: knowns and unknowns* (S. 127—150). New York, NY: Oxford University Press.
- Jäger, A. O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M. & Beauducel, A. (2006). *BIS-HB: Berliner Intelligenzstrukturtest für Jugendliche: Begabungs- und Hochbegabungsdiagnostik*. Göttingen: Hogrefe.
- James, W. (1884). What is an emotion? *Mind*, *9*, 188—205. doi: 10.1093/mind/os-IX.34.188
- Jensen, A. R. (1992). Understanding *g* in terms of information processing. *Educational Psychology Review*, *4*, 271—308. doi: 10.1007/BF01417874
- John, O. P., Naumann, L. P. & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy: history, measurement, and conceptual issues. In O. P. John, R. W. Robins & L. A. Pervin (Eds.), *Handbook of personality: theory and research* (3. Aufl., S. 114—158). New York, NY: The Guildford Press.

- John, O. P. & Srivastava, S. (1999). The big five trait taxonomy: history, measurement, and theoretical perspectives. In L. Pervin & O. P. John (Eds.), *Handbook of personality: theory and research* (2. Aufl., S. 102—138). New York, NY: The Guildford Press.
- Joseph, D. L. & Newman, D. A. (2010). Emotional intelligence: an integrative meta-analysis and cascading model. *Journal of Applied Psychology*, *95*, 54—78. doi: 10.1037/a0017286
- Kammrath, L. K., Mendoza-Denton, R. & Mischel, W. (2005). Incorporating if... then... personality signatures in person perception: beyond the person-situation dichotomy. *Journal of Personality and Social Psychology*, *88*, 605—618. doi: 10.1037/0022-3514.88.4.605
- Kang, S.-M., Day, J. D. & Meara, N. M. (2005). Social and emotional intelligence: Starting a conversation about their similarities and differences. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: an international handbook* (S. 91—105). Cambridge, MA: Hogrefe & Huber.
- Kee, K. S., Horan, W. P., Salovey, P., Kern, R. S., Sergi, M. J., Fiske, A. P., ... Green, M. F. (2009). Emotional intelligence in schizophrenia. *Schizophrenia Research*, *107*, 61—68. doi: 10.1016/j.schres.2008.08.016
- Keefer, K. V., Parker, J. D. A. & Saklofske, D. H. (2009). Emotional Intelligence and physical health. In C. Stough, D. H. Saklofske & J. D. A. Parker (Eds.), *Assessing emotional intelligence: theory, research, and application* (S. 191—218). New York, NY: Springer.
- Keele, S. M. & Bell, R. C. (2009). Consensus scoring, correct responses and reliability of the MSCEIT V2. *Personality and Individual Differences*, *47*, 740—747. doi: 10.1016/j.paid.2009.06.013



- Kenny, D. A., Kashy, D. A. & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: The Guilford Press.
- Klein, P. D. (1997). Multiplying the problems of intelligence by eight: a critique of Gardner's theory. *Canadian Journal of Education*, 22, 377—394. doi: 10.2307/1585790
- Kring, A. M., Smith, D. A. & Neale, J. M. (1994). Individual differences in dispositional expressiveness: Development and validation of the Emotional Expressivity Scale. *Journal of Personality and Social Psychology*, 66, 934—949. doi: 10.1037/0022-3514.66.5.934
- Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77, 1121—1134. doi: 10.1037/0022-3514.77.6.1121
- Lamm, C., Meltzoff, A. N. & Decety, J. (2010). How do we empathize with someone who is not like us? A functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*, 22, 362—376. doi: 10.1162/jocn.2009.21186
- Landy, F. J. (2006). The long, frustrating, and fruitless search for social intelligence: a cautionary tale. In K. R. Murphy (Ed.), *A critique of emotional intelligence: What are the problems and how can they be fixed?* (S. 81—123). New York, NY: Lawrence Erlbaum Associates.
- Lane, R. D., Quinlan, D. M., Schwartz, G. E., Walker, P. A. & Zeitlin, S. B. (1990). The Levels of Emotional Awareness Scale: a cognitive-developmental measure of emotion. *Journal of Personality Assessment*, 55, 124—134. doi: 10.1080/00223891.1990.9674052

- Lee, G., Brennan, R. L. & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practice*, 19, 9—15. doi: 10.1111/j.1745-3992.2000.tb00041.x
- Lee, G., Dunbar, S. B. & Frisbie, D. A. (2001). The relative appropriateness of eight measurement models for analyzing scores from tests composed of testlets. *Educational and Psychological Measurement*, 61, 958—975. doi: 10.1177/00131640121971590
- Legree, P. J., Psozka, J., Tremble, T. & Bourne, D. R. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: an international handbook* (S. 156—179). Cambridge, MA: Hogrefe & Huber.
- Leuner, B. (1966). Emotionale Intelligenz und Emanzipation: Eine psychodynamische Studie über die Frau. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, 15(6), 196—203.
- Libbrecht, N. & Lievens, F. (2012). Validity evidence for the situational judgment test paradigm in emotional intelligence measurement. *International Journal of Psychology*, 47, 438—447. doi: 10.1080/00207594.2012.682063
- Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2006). *Intelligenz-Struktur-Test 2000 R* (2. Aufl.). Göttingen: Hogrefe.
- Lievens, F., Klehe U.-C. & Libbrecht, N. (2011). Applicant versus employee scores on self-report emotional intelligence measures. *Journal of Personnel Psychology*, 10, 89—95. doi: 10.1027/1866-5888/a000036
- Little, T. D., Cunningham, W. A., Shahar, G. & Widaman, K. F. (2002). To parcel or not to parcel: exploring the question, weighing the merits. *Structural Equation Modeling*, 9, 151—173. doi: 10.1207/S15328007SEM0902\_1

- Livingstone, H. A. & Day, A. L. (2005). Comparing the construct and criterion-related validity of ability-based and mixed-model measures of emotional intelligence. *Educational and Psychological Measurement, 65*, 757—779. doi: 10.1177/0013164405275663
- Lopes, P. N., Mestre, J. M., Guil, R., Kremenitzer, J. P. & Salovey, P. (2012). The role of knowledge and skills for managing emotions in adaptation to school: social behavior and misconduct in the classroom. *American Educational Research Journal, 49*, 710—742. doi: 10.3102/0002831212443077
- Lord, F. M. & Novick, M. R. (1974). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing Company.
- Losoya, S. H. & Eisenberg, N. (2001). Affective empathy. In J. A. Hall & F. Bernieri (Eds.), *Interpersonal sensitivity: theory and measurement* (S. 21—43). Mahwah, NJ: Lawrence Erlbaum Associates.
- MacCann, C. (2010). Further examination of emotional intelligence as a standard intelligence: A latent variable analysis of fluid intelligence, crystallized intelligence, and emotional intelligence. *Personality and Individual Differences, 49*, 490—496. doi: 10.1016/j.paid.2010.05.010
- MacCann, C., Pearce, N. & Roberts, R. D. (2011). Emotional intelligence as assessed by situational judgment and emotion recognition tests: building the nomological net. *Psihologijske Teme, 20*(3), 393—412.
- MacCann, C. & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: theory and data. *Emotion, 8*, 540—551. doi: 10.1037/a0012746
- MacCann, C., Roberts, R. D., Matthews, G. & Zeidner, M. (2004). Consensus scoring and empirical option weighting of performance-based Emotional Intelligence (EI) tests.

- Personality and Individual Differences*, 36, 645—662. doi: 10.1016/S0191-8869(03)00123-5
- Matsumoto, D., Keltner, D., Shiota, M. N., O'Sullivan, M. & Frank, M. (2008). Facial expressions of emotion. In M. Lewis, J. M. Haviland-Jones, and L. Feldman Barrett (Eds.), *Handbook of emotions* (3. Aufl., S. 211—234). New York, NY: The Guilford Press.
- Matsumoto, D., LeRoux, J., Wilson-Cohn, C., Raroque, J., Kookan, K., Ekman, P., ... Goh, A. (2000). A new test to measure emotion recognition ability: Matsumoto and Ekman's Japanese and Caucasian Brief Affect Recognition Test (JACBART). *Journal of Nonverbal Behavior*, 24, 179—209. doi: 10.1023/A:1006668120583
- Matthews, G., Zeidner, M. & Roberts, R. D. (2002). *Emotional intelligence: science and myth*. Cambridge, MA: MIT Press.
- Maul, A. (2012). Examining the structure of emotional intelligence at the item level: new perspectives, new conclusions. *Cognition & Emotion*, 26, 503—520. doi: 10.1080/02699931.2011.588690
- Mayer, J. D. (2007-2008). Personal intelligence. *Imagination, Cognition and Personality*, 27, 209—232. doi: 10.2190/IC.27.3.b
- Mayer, J. D., Caruso, D. R. & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, 27, 267—298. doi: 10.1016/S0160-2896(99)00016-1
- Mayer, J. D. & Geher, G. (1996). EI and the identification of emotion. *Intelligence*, 22, 89—113. doi: 10.1016/S0160-2896(96)90011-2

- Mayer, J. D., Roberts, R. D. & Barsade, S. G. (2008). Human abilities: emotional intelligence. *Annual Review of Psychology*, *59*, 507—536. doi: 10.1146/annurev.psych.59.103006.093646
- Mayer, J. D. & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. Sluyter (Eds.), *Emotional development and emotional intelligence: implications for educators* (S. 3—31). New York, NY: Basic Books.
- Mayer, J. D., Salovey, P. & Caruso, D. (2000). Models of emotional intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (S. 396—420). Cambridge, UK: Cambridge University Press.
- Mayer, J. D., Salovey, P. & Caruso, D. R. (2002). *Mayer-Salovey-Caruso Emotional Intelligence Test MSCEIT: user's manual*. Toronto, Kanada: Multi-Health Systems.
- Mayer, J. D., Salovey, P., Caruso, D. R. & Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion*, *1*, 232—242. doi: 10.1037/1528-3542.1.3.232
- Mayer, J. D., Salovey, P., Caruso, D. R. & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion*, *3*, 97—105. doi: 10.1037/1528-3542.3.1.97
- McCrae, R. R. & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*, 175—215. doi: 10.1111/j.1467-6494.1992.tb00970.x
- McDaniel, M. A. & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, *9*, 103—113. doi: 10.1111/1468-2389.00167

- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741—749. doi: 10.1037/0003-066X.50.9.741
- Meyvis, T., Ratner, R. K. & Levav, J. (2010). Why don't we learn to accurately forecast feelings? How misremembering our predictions blinds us to past forecasting errors. *Journal of Experimental Psychology: General*, *139*, 579—589. doi: 10.1037/a0020285
- Michaels, T. M., Horan, W. P., Ginger, E. J., Martinovich, Z., Pinkham, A. E. & Smith, M. J. (2014). Cognitive empathy contributes to poor social functioning in schizophrenia: Evidence from a new self-report measure of cognitive and affective empathy. *Psychiatry Research*, *220*, 803—810. doi: 10.1016/j.psychres.2014.08.054
- Mikolajczak, M., Nelis, D., Hansenne, M. & Quoidbach, J. (2008). If you can regulate sadness, you can probably regulate shame: Associations between trait emotional intelligence, emotion regulation and coping efficiency across discrete emotions. *Personality and Individual Differences*, *44*, 1356—1368. doi: 10.1016/j.paid.2007.12.004
- Mohorić, T., Takšić, V. & Duran, M. (2010). In search of “the correct answer” in an ability-based Emotional Intelligence (EI) test. *Studia Psychologica*, *52*(3), 219—228.
- Moors, A. (2009). Theories of emotion causation: a review. *Cognition & Emotion*, *23*, 625—662. doi: 10.1080/02699930802645739
- Muthén, L. K. & Muthén, B. O. (2010). *Mplus user's guide* (6. Aufl.). Los Angeles, CA: Muthén & Muthén.

- Nelis, D., Quoidbach, J., Mikolajczak, M. & Hansenne, M. (2009). Increasing emotional intelligence: (How) is it possible? *Personality and Individual Differences*, *47*, 36—41. doi: 10.1016/j.paid.2009.01.046
- Neubauer, A. C. & Freudenthaler, H. H. (2005). Models of emotional intelligence. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: an international handbook* (S. 31—50). Cambridge, MA: Hogrefe & Huber.
- Niedenthal, P. M., Krauth-Gruber, S. & Ric, F. (2006). *Psychology of emotion: interpersonal, experiential, and cognitive approaches*. New York, NY: Psychology Press.
- Nowicki, S., Jr. & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: the Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior*, *18*, 9—35. doi: 10.1007/BF02169077
- O'Connor, R. M., Jr. & Little I. S. (2003). Revisiting the predictive validity of emotional intelligence: self-report versus ability-based measures. *Personality and Individual Differences*, *35*, 1893—1902. doi: 10.1016/S0191-8869(03)00038-2
- Olsen, J. A. & Kenny, D. A. (2006). Structural equation modeling with interchangeable dyads. *Psychological Methods*, *11*, 127—141. doi: 10.1037/1082-989X.11.2.127
- Orchard, B., MacCann, C., Schulze, R., Matthews, G., Zeidner, M. & Roberts, R. D. (2009). New directions and alternative approaches to the measurement of emotional intelligence. In C. Stough, D. H. Saklofske & J. D. A. Parker (Eds.), *Assessing emotional intelligence: theory, research, and application* (S. 321—344). New York, NY: Springer.
- Ostendorf, F. & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae* (Rev. Aufl.). Göttingen: Hogrefe.

- Oswald, P. A. (1996). The effects of cognitive and affective perspective taking on empathic concern and altruistic helping. *The Journal of Social Psychology, 136*, 613—623. doi: 10.1080/00224545.1996.9714045
- O`Sullivan, M. & Ekman, P. (2004). Facial expression recognition and emotional intelligence. In G. Geher (Ed.), *Measuring emotional intelligence: common ground and controversy* (S. 91—111). Hauppauge, NY: Nova Science Publishers.
- O`Sullivan, M. & Guilford, J. P. (1975). Six factors of behavioral cognition: understanding other people. *Journal of Educational Measurement, 12*, 255—271. doi: 10.1111/j.1745-3984.1975.tb01027.x
- Palfai, T. P. & Salovey, P. (1993). The influence of depressed and elated mood on deductive and inductive reasoning. *Imagination, Cognition and Personality, 13*, 57—71. doi: 10.2190/FYYA-GCRU-J124-Q3B2
- Palmer, B. R., Gignac, G., Manocha, R. & Stough, C. (2005). A psychometric evaluation of the Mayer–Salovey–Caruso Emotional Intelligence Test Version 2.0. *Intelligence, 33*, 285—305. doi: 10.1016/j.intell.2004.11.003
- Pérez J. C., Petrides, K. V. & Furnham, A. (2005). Measuring trait emotional intelligence. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: an international handbook* (S. 181—201). Cambridge, MA: Hogrefe & Huber.
- Petrides, K. V. & Furnham, A. (2000). On the dimensional structure of emotional intelligence. *Personality and Individual Differences, 29*, 313—320. doi: 10.1016/S0191-8869(99)00195-6



- Petrides, K. V. & Furnham, A. (2001). Trait emotional intelligence: Psychometric investigation with reference to established trait taxonomies. *European Journal of Personality, 15*, 425—448. doi: 10.1002/per.416
- Petrides, K. V. & Furnham, A. (2003). Trait emotional intelligence: Behavioural validation in two studies of emotion recognition and reactivity to mood induction. *European Journal of Personality, 17*, 39—57. doi: 10.1002/per.466
- Petrides, K. V., Furnham, A. & Mavroveli, S. (2007). Trait emotional intelligence: Moving forward in the field of EI. In G. Matthews, M. Zeidner & R. D. Roberts (Eds.), *The science of emotional intelligence: knowns and unknowns* (S. 151—166). New York, NY: Oxford University Press.
- Petrides, K. V., Pita, R. & Kokkinaki, F. (2007). The location of trait emotional intelligence in personality factor space. *British Journal of Psychology, 98*, 273—289. doi: 10.1348/000712606X120618
- Peugh, J. L., DiLillo, D. & Panuzio, J. (2013). Analyzing mixed-dyadic data using structural equation models. *Structural Equation Modeling, 20*, 314—337. doi: 10.1080/10705511.2013.769395
- Pollmann, M. M. H. & Finkenauer, C. (2009). Empathic forecasting: How do we predict other people's feelings? *Cognition and Emotion, 23*, 978—1001. doi: 10.1080/02699930802264895
- Pons, F., Harris, P. L. & de Rosnay, M. (2004). Emotion comprehension between 3 and 11 years: developmental periods and hierarchical organization. *European Journal of Developmental Psychology, 1*, 127—152. doi: 10.1080/17405620344000022

- Qualter, P., Barlow, A. & Stylianou, M. S. (2011). Investigating the relationship between trait and ability emotional intelligence and theory of mind. *British Journal of Developmental Psychology*, 29, 437—454. doi: 10.1348/026151010X502999
- Riggio, R. E., Messamer, J. & Throckmorton, B. (1991). Social and academic intelligence: conceptually distinct but overlapping constructs. *Personality and Individual Differences*, 12, 695—702. doi: 10.1016/0191-8869(91)90225-Z
- Rimé, B., Mesquita, B., Philippot, P. & Boca, S. (1991). Beyond the emotional event: Six studies on the social sharing of emotion. *Cognition & Emotion*, 5, 435—465. doi: 10.1080/02699939108411052
- Rivers, S. E., Brackett, M. A., Reyes, M. R., Mayer, J. D., Caruso, D. R. & Salovey, P. (2012). Measuring emotional intelligence in early adolescence with the MSCEIT-YV: psychometric properties and relationship with academic performance and psychosocial functioning. *Journal of Psychoeducational Assessment*, 30, 344—366. doi: 10.1177/0734282912449443
- Roberts, R. D., Schulze, R. & MacCann, C. (2008). The measurement of emotional intelligence: a decade of progress? In G. Boyle, G. Matthews & D. Saklofske (Eds.), *The Sage handbook of personality theory and assessment* (S. 461—482). New York, NY: Sage.
- Roberts, R. D., Schulze, R., O'Brien, K., MacCann, C., Reid, J. & Maul, A. (2006). Exploring the validity of the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) with established emotions measures. *Emotion*, 6, 663—669. doi: 10.1037/1528-3542.6.4.663

- Roberts, R. D., Zeidner M. & Matthews, G. (2001). Does emotional intelligence meet traditional standards for an intelligence? Some new data and conclusions. *Emotion, 1*, 196—231. doi: 10.1037/1528-3542.1.3.196
- Robitzsch, A. (2014). Supplementary Item Response Theory Model (sirt) [Computer software]. Retrieved from <https://sites.google.com/site/alexanderrobitzsch/software>
- Roseman, I. J. (1991). Appraisal determinants of discrete emotions. *Cognition & Emotion, 5*, 161—200. doi: 10.1080/02699939108411034
- Roseman, I. J. (2001). A model of appraisal in the emotion system: integrating theory, research, and applications. In K. R. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: theory, methods, research* (S. 68—91). New York, NY: Oxford University Press.
- Roseman, I. J. & Evdokas, A. (2004). Appraisals cause experienced emotions: experimental evidence. *Cognition & Emotion, 18*, 1—28. doi: 10.1080/02699930244000390
- Roseman, I. J. & Smith, G. (2001). Appraisal theory: overview, assumptions, varieties, controversies. In K. R. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: theory, methods, research* (S. 3—19). New York, NY: Oxford University Press.
- Roseman, I. J., Spindel, M. S. & Jose, P. E. (1990). Appraisals of emotion-eliciting events: Testing a theory of discrete emotions. *Journal of Personality and Social Psychology, 59*, 899—915. doi: 10.1037/0022-3514.59.5.899
- Rossen, E. & Kranzler, J. H. (2009). Incremental validity of the Mayer–Salovey–Caruso Emotional Intelligence Test Version 2.0 (MSCEIT) after controlling for personality

- and intelligence. *Journal of Research in Personality*, 43, 60—65. doi: 10.1016/j.jrp.2008.12.002
- Rossen, E., Kranzler, J. H. & Algina, J. (2008). Confirmatory factor analysis of the Mayer–Salovey–Caruso Emotional Intelligence Test V 2.0 (MSCEIT). *Personality and Individual Differences*, 44, 1258—1269. doi: 10.1016/j.paid.2007.11.020
- Rost, J. (2004). *Testtheorie-Testkonstruktion* (2. rev. Aufl.). Bern: Hans Huber.
- Ruby, P. & Decety, J. (2004). How would you feel versus how do you think she would feel? A neuroimaging study of perspective-taking with social emotions. *Journal of Cognitive Neuroscience*, 16, 988—999. doi: 10.1162/0898929041502661
- Salovey, P. & Mayer, J. D. (1990). Emotional Intelligence. *Imagination, Cognition and Personality*, 9(3), 185—211.
- Schachter, S. & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, 69, 379—399. doi: 10.1037/h0046234
- Scherer, K. R. (1993). Studying the emotion-antecedent appraisal process: an expert system approach. *Cognition & Emotion*, 7, 325—355. doi: 10.1080/02699939308409192
- Scherer, K. R. (1999). Appraisal theory. In T. Dalgleish & M. J. Power (Eds.), *Handbook of cognition and emotion* (S. 637—663). New York, NY: John Wiley & Sons.
- Scherer, K. R. & Scherer, U. (2011). Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the Emotion Recognition Index. *Journal of Nonverbal Behavior*, 35, 305—326. doi: 10.1007/s10919-011-0115-4

- Schlegel, K., Grandjean, D. & Scherer, K. R. (2012). Emotion recognition: Unidimensional ability or a set of modality- and emotion-specific skills? *Personality and Individual Differences*, *53*, 16—21. doi: 10.1016/j.paid.2012.01.026
- Schulte, M. J., Ree, M. J. & Carretta T. R. (2004). Emotional intelligence: not much more than *g* and personality. *Personality and Individual Differences*, *37*, 1059—1068. doi: 10.1016/j.paid.2003.11.014
- Schulte-Rüther, M., Markowitsch, H. J., Fink, G. R. & Piefke, M. (2007). Mirror neuron and theory of mind mechanisms involved in face-to-face interactions: a functional magnetic resonance imaging approach to empathy. *Journal of Cognitive Neuroscience*, *19*, 1354—1372. doi: 10.1162/jocn.2007.19.8.1354
- Schulze, R., Hellwig, S., MacCann, C. & Roberts, R. D. (2010, September). *Die Erfassung von Emotional Understanding mit dem Empathic Agent Paradigma*. Vortrag gehalten auf dem 47. Kongress der Deutschen Gesellschaft für Psychologie, Bremen.
- Schulze, R., Holtzman, S., MacCann, C. & Roberts, R. D. (2009, September). *Assessment of emotional understanding with the Empathic Agent Paradigm (EAP)*. Vortrag gehalten auf der zehnten European Conference on Psychological Assessment, Ghent, Belgien.
- Schulze, R. & Roberts, R. D. (2006). Assessing the big five: Development and validation of the Openness Conscientiousness Extraversion Agreeableness Neuroticism Index Condensed (OCEANIC). *Zeitschrift für Psychologie*, *214*, 133—149. doi: 10.1026/0044-3409.214.3.133
- Schulze, R., Roberts, R. D., Zeidner, M. & Matthews, G. (2005). Theory, measurement, and applications of emotional intelligence: frames of reference. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: an international handbook* (S. 3—29). Cambridge, MA: Hogrefe & Huber.

- Schulze, R., Wilhelm, O. & Kyllonen, P. C. (2007). Approaches to the assessment of emotional intelligence. In G. Matthews, M. Zeidner & R. D. Roberts (Eds.), *The science of emotional intelligence: knowns and unknowns* (S. 199—229). New York, NY: Oxford University Press.
- Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J. & Dornheim, L. (1998). Development and validation of a measure of emotional intelligence. *Personality and Individual Differences*, 25, 167—177. doi: 10.1016/S0191-8869(98)00001-4
- Schutte, N. S., Malouff, J. M., Thorsteinsson, E. B., Bhullar, N. & Rooke, S. E. (2007). A meta-analytic investigation of the relationship between emotional intelligence and health. *Personality and Individual Differences*, 42, 921—933. doi: 10.1016/j.paid.2006.09.003
- Shaver, P., Schwartz, J., Kirson, D. & O'Connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52, 1061—1086. doi: 10.1037/0022-3514.52.6.1061
- Sheldon, O. J., Dunning, D. & Ames, D. R. (2014). Emotionally unskilled, unaware, and uninterested in learning more: Reactions to feedback about deficits in emotional intelligence. *Journal of Applied Psychology*, 99, 125—137. doi: 10.1037/a0034138
- Sireci, S. G., Thissen, D. & Wainer, H. (1991). *On the reliability of testlet-based tests* (Technical Report No. 91-10). Princeton, NJ: Educational Testing Service.
- Spearman, C. (1920). Manifold sub-theories of “the two factors”. *Psychological Review*, 27, 159—172. doi: 10.1037/h0068562

- Spreng, R. N., McKinnon, M. C., Mar, R. A. & Levine, B. (2009). The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment*, *91*, 62—71. doi: 10.1080/00223890802484381
- Steinmayr, R., Schütz, A., Hertel J. & Schröder-Abé M. (2011). *Mayer-Salovey-Caruso Test zur Emotionalen Intelligenz (MSCEIT). Deutschsprachige Adaptation des Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) von John. D. Mayer, Peter Salovey und David R. Caruso*. Bern: Hans Huber.
- Süß, H.-M. & Beauducel, A. (2005). Faceted models of intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (S. 313—332). Thousand Oaks, CA: Sage Publications.
- Thorndike, E. L. (1920, Januar). Intelligence and its use. *Harper's Magazine*, *140*(1), 227—235.
- Trentacosta, C. J. & Izard, C. E. (2007). Kindergarten children's emotion competence as a predictor of their academic competence in first grade. *Emotion*, *7*, 77—88. doi: 10.1037/1528-3542.7.1.77
- Vachon, D. D., Lynam, D. R. & Johnson, J. A. (2014). The (non)relation between empathy and aggression: surprising results from a meta-analysis. *Psychological Bulletin*, *140*, 751—773. doi: 10.1037/a0035236
- Van Boven, L. & Loewenstein, G. (2005). Empathy gaps in emotional perspective taking. In B. F. Malle & S. D. Hodges (Eds.), *Other minds: How humans bridge the divide between self and others* (S. 284—297). New York, NY: The Guilford Press.

- Van Dijk, W. W., Van Dillen, L. F., Seip, E. C. & Rotteveel, M. (2012). Emotional time travel: Emotion regulation and the overestimation of future anger and sadness. *European Journal of Social Psychology, 42*, 308—313. doi: 10.1002/ejsp.1853
- Van Rooy, D. L., Viswesvaran, C. & Pluta, P. (2005). An evaluation of construct validity: What is this thing called emotional intelligence? *Human Performance, 18*, 445—462. doi: 10.1207/s15327043hup1804\_9
- von Stumm, S. & Ackerman, P. L. (2013). Investment and intellect: a review and meta-analysis. *Psychological Bulletin, 139*, 841—869. doi: 10.1037/a0030746
- Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.
- Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: a case for testlets. *Journal of Educational Measurement, 24*(3), 185—201.
- Wang, X., Bradlow, E. T. & Wainer, H. (2002). A general Bayesian model for testlets: theory and applications. *Applied Psychological Measurement, 26*, 109—128. doi: 10.1177/0146621602026001007
- Waterhouse, L. (2006). Multiple intelligences, the Mozart effect, and emotional intelligence: a critical review. *Educational Psychologist, 41*, 207—225. doi: 10.1207/s15326985ep4104\_1
- Watson, J. B. & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology, 3*, 1—14. doi: 10.1037/h0069608
- Webb, T. L., Miles, E. & Sheeran, P. (2012). Dealing with feeling: A meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation. *Psychological Bulletin, 138*, 775—808. doi: 10.1037/a0027600



- Weekley, J. A. & Ployhart, R. E. (2006). An introduction to situational judgment testing. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment test: theory, measurement, and application* (S. 1—10). Mahwah, NJ: Lawrence Erlbaum Associates.
- 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 (Eds.), *Situational judgment test: theory, measurement, and application* (S. 157—182). Mahwah, NJ: Lawrence Erlbaum Associates.
- Weis, S. & Süß, H.-M. (2005). Social intelligence – A review and critical discussion of measurement concepts. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: an international handbook* (S. 203—230). Cambridge, MA: Hogrefe & Huber.
- Weis, S. & Süß, H.-M. (2007). Reviving the search for social intelligence – A multitrait-multimethod study of its structure and construct validity. *Personality and Individual Differences*, 42, 3—14. doi: 10.1016/j.paid.2006.04.027
- Wilhelm, O. (2005). Measures of emotional intelligence: practice and standards. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: an international handbook* (S. 131—154). Cambridge, MA: Hogrefe & Huber.
- Wilhelm, O., Schroeders, U. & Schipolowski, S. (2014). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. Jahrgangsstufe (BEFKI 8-10)*. Göttingen: Hogrefe.
- Wilson, T. D. & Gilbert, D. T. (2003). Affective forecasting. *Advances in Experimental Social Psychology*, 35, 345—411. doi: 10.1016/S0065-2601(03)01006-2

- Wilson, T. D., Meyers, J. & Gilbert, D. T. (2001). Lessons from the past: Do people learn from experience that emotional reactions are short-lived? *Personality and Social Psychology Bulletin*, 27, 1648—1661. doi: 10.1177/01461672012712008
- Wong, C.-S. & Law, K. S. (2002). The effects of leader and follower emotional intelligence on performance and attitude: an exploratory study. *The Leadership Quarterly*, 13, 243—274. doi: 10.1016/S1048-9843(02)00099-1
- Wranik, T., Feldman Barrett, L. & Salovey, P. (2007). Intelligent emotion regulation: Is knowledge power? In J. J. Gross (Ed.), *Handbook of emotion regulation* (S. 393—407). New York, NY: The Guilford Press.
- Zaki, J., Bolger, N. & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science*, 19, 399—404. doi: 10.1111/j.1467-9280.2008.02099.x
- Zeidner, M., Matthews, G. & Roberts, R. D. (2001). Slow down, you move too fast: Emotional intelligence remains an “elusive” intelligence. *Emotion*, 1, 265—275. doi: 10.1037/1528-3542.1.3.265
- Zeidner, M., Roberts, R. D. & Matthews, G. (2008). The science of emotional intelligence: current consensus and controversies. *European Psychologist*, 13, 64—78. doi: 10.1027/1016-9040.13.1.64