

Schulerfolg, Lebenszufriedenheit und psychische Gesundheit:
Differentialpsychologische Untersuchungen zu Fragen der (subjektiven)
Lebensqualität

Dissertation

zur Erlangung des akademischen Grades eines

Doktors der Philosophie (Dr. phil.)

der Fakultät HW

Bereich Empirische Humanwissenschaften

der Universität des Saarlandes

vorgelegt von

Alexander Dings

aus Saarbrücken

Saarbrücken, 2026

Dekan: Prof. Dr. Axel Mecklinger, Universität des Saarlandes

1. Berichterstatter: Prof. Dr. Frank M. Spinath, Universität des Saarlandes

2. Berichterstatter: Prof. Dr. Cornelius J. König, Universität des Saarlandes

Tag der Disputation: 15. 04. 2026

Inhaltsverzeichnis

Abkürzungsverzeichnis.....	V
Tabellenverzeichnis.....	VI
Abstract.....	1
Relevante Studien.....	3
1 Einleitung.....	4
TwinLife.....	4
2 Schulisches Underachievement.....	6
Definition und Operationalisierung.....	6
Das Achievement Orientation Model (AOM).....	9
Ergebnisse und Schlussfolgerungen.....	11
3 Twin-Difference-Analyse der Zusammenhänge zwischen subjektiver Lebenszufriedenheit und dem Besuch von Sportvereinen.....	18
Methoden der Zwillingsforschung und das missing environment.....	18
Subjektive Lebenszufriedenheit.....	20
Sportliche Aktivität und Lebenszufriedenheit.....	21
Ergebnisse und Schlussfolgerungen.....	22
4 Labeling-Effekte im Kontext psychischer Erkrankungen.....	29
Diagnostische Label als mögliche Treiber von Stigmatisierung.....	29
Labeling-Effekte auf der Ebene einzelner Verhaltensweisen.....	31
Methoden.....	32
Ergebnisse und Schlussfolgerungen.....	34
5 Zusammenfassung.....	41
Literaturverzeichnis.....	44
Danksagung	62
Anhang.....	63

Abkürzungsverzeichnis

ADE-Modell	univariates genetisches Modell mit Berücksichtigung von Dominanzeffekten (D)
AOM	Achievement Orientation Model
ASKU	Allgemeine Selbstwirksamkeit Kurzsкала
β_B	between-pair effect
β_W	within-pair effect
BFI-S	Big Five Inventory-SOEP
BIC	Bayesian Information Criterion
CFI	Comparative Fit Index
CFT	Culture Fair Test
DLS	Domänenspezifische Lebenszufriedenheit
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, 5. Auflage
DZ	dizygot (zweieiig)
GLS	Globale Lebenszufriedenheit
M	Mittelwert
MZ	monozygot (eineiig)
RMSEA	Root Mean Square Error of Approximation
SD	Standardabweichung
SELLMO	Skalen zur Erfassung der Lern- und Leistungsmotivation
SWLS	Satisfaction with Life Scale
SNP	Single Nucleotide Polymorphism
SOEP	Sozioökonomisches Panel
SWB	Subjektives Wohlbefinden

Tabellenverzeichnis

Tabelle 1 Einteilung der Schüler nach der <i>absolute split method</i> in Studie I.....	7
Tabelle 2 Mittelwerte und Standardabweichungen der Prädiktorvariablen in den für die Studie definierten Gruppen.....	13
Tabelle 3 Geschätzte Varianzanteile der Lebenszufriedenheits-Maße in univariaten Zwillingsmodellen.....	24
Tabelle 4 Regressionsgewichte aus linearen gemischten Modellen zur Vorhersage der globalen und domänenspezifischen Lebenszufriedenheit aus Unterschieden in der Teilnahme an Sportvereinen innerhalb und zwischen Zwillingspaaren sowie Alter und Geschlecht.....	25
Tabelle 5 Fixed effects aus den linearen Modellen mit der Bewertung einer Verhaltensweise als Outcome.....	36
Tabelle 6 Modellfitstatistiken zu konfirmatorischen Faktorenanalysen getrennt nach Bewertungskategorie und Experimentalgruppe.....	38

Abstract

Die vorliegende kumulative Dissertation befasst sich im Rahmen verschiedener inhaltlicher Domänen und methodischer Ansätze mit Aspekten der subjektiven Lebensqualität. Dabei stützen sich zwei der vorgestellten Studien auf das Zwillingsprojekt *TwinLife*, das rund 4000 Zwillingspaare und ihre Familien in Deutschland während des Heranwachsens und jungen Erwachsenenalters begleitet. In dieser Panelstudie stehen Variablen im Fokus, die es ermöglichen, unterschiedliche Lebenswege nachzuvollziehen und zu ergründen, wie es zu Unterschieden in der Lebensqualität, dem beruflichen und privaten Erfolg und zu sozialer Ungleichheit kommt (siehe Abschnitt 1). Die Studien I und II beleuchten zwei Teilaspekte dieser Thematik - schulischen Erfolg sowie Lebenszufriedenheit – während Studie III die kumulative Dissertation um die Untersuchung einer anderen Facette der subjektiven Lebensqualität und der sozialen Integration außerhalb des TwinLife-Panels ergänzt, nämlich der Stigmatisierung psychischer Erkrankungen.

Studie I (Dings & Spinath, 2021), die in Abschnitt 2 geschildert wird, setzt sich mit der Frage auseinander, unter welchen Bedingungen Schüler¹ trotz guter allgemeiner Intelligenz keine guten Schulleistungen erreichen und was dieses sogenannte *Underachievement* von anderen Fällen schulischen Misserfolgs abgrenzt. Mithilfe einer Diskriminanzanalyse zeigt Studie I, dass erfolglose Schüler eine geringere schulische Motivation und elterliche Unterstützung berichten als erfolgreiche Schüler. Über diesen Vergleich hinaus, der den Erkenntnissen früherer Forschung entspricht, ließen sich Underachiever in einem in dieser Form neuartigen Ergebnis zudem auch von anderen erfolglosen Schülern abgrenzen. Angesichts einer geringen Stichprobengröße nur vorsichtig zu interpretierende Ergebnisse deuten hier auf eine bisher möglicherweise unterschätzte Bedeutung von Persönlichkeitsfaktoren statt motivationalen Variablen hin.

Studie II (Dings & Spinath, 2024) untersucht mögliche Zusammenhänge des regelmäßigen Besuchs von Sportvereinen und der Lebenszufriedenheit. Dabei wird die Zwillingsstichprobe des TwinLife-Panels genutzt, um zu zeigen, dass eine höhere domänenspezifische Zufriedenheit mit der eigenen Gesundheit und Freizeitgestaltung bei Zwillingen, die regelmäßig einen Sportverein besuchen, nicht durch eine mögliche genetische und geteilt umweltbedingte Konfundierung erklärbar ist. Den Darstellungen der Ergebnisse dieser Arbeit in Abschnitt 3 gehen allgemeine Einführungen zu verhaltensgenetischen Methoden und ihrer Relevanz für Studie II voraus.

Abschnitt 4 der Dissertation befasst sich außerhalb von TwinLife mit einem anderen Aspekt der subjektiven Lebensqualität, indem in Studie III (Dings & Kubillus, 2025) sogenannte Labeling-

¹ Die vorliegende Abhandlung nutzt aus Gründen der Lesbarkeit das generische Maskulinum und meint damit – wenn nicht anders gekennzeichnet – alle Geschlechter.

Effekte im Kontext psychischer Erkrankungen untersucht werden. In einem neuartigen methodischen Ansatz wird hier gezeigt, dass Symptome psychischer Erkrankungen bei Erwähnung eines diagnostischen Labels teilweise anders bewertet werden als ohne eine solche Etikettierung. In allen vorgestellten Studien liegt ein Fokus auf dem Versuch, die bisherigen Forschungsergebnisse zur jeweiligen Fragestellung durch methodische Innovationen zu ergänzen. In der folgenden Synopse wird deshalb vor der jeweiligen Ergebnisdarstellung auf die bisherigen methodischen Herausforderungen und die in den Studien I bis III dafür gefundenen Lösungen eingegangen.

Relevante Studien

Die vorliegende kumulative Dissertation umfasst die folgenden drei Studien, die in internationalen wissenschaftlichen Zeitschriften nach einem Peer-Review-Prozess veröffentlicht wurden. Im Falle von Studie III wurde nach bereits erfolgter Annahme und Publikation die Notwendigkeit einer inhaltlichen Korrektur festgestellt. Ein entsprechendes „Erratum“ wurde zum Zeitpunkt der vorliegenden Arbeit aufseiten des betreffenden Journals bereits zur Veröffentlichung unter voraussichtlich derselben Zitation angenommen, eine tatsächliche Publikation steht jedoch noch aus.

Die Manuskripte der Studien I und II sind im Anhang dieser Arbeit in ihrer vollständigen publizierten Form aufgeführt. Für Studie III wird dort das bereits akzeptierte, aber noch nicht in dieser Form veröffentlichte korrekte Manuskript aufgeführt.

Studie I

Dings, A., & Spinath, F. M. (2021). Motivational and personality variables distinguish academic underachievers from high achievers, low achievers, and overachievers. *Social Psychology of Education*, 24(6), 1461-1485. <https://doi.org/10.1007/s11218-021-09659-2>

Zur Veröffentlichung akzeptiert am 10. 8. 2021.

Studie II

Dings, A., & Spinath, F. M. (2024). Sports club participation impacts life satisfaction in adolescence: A twin study. *Psychology of Sport and Exercise*, 73, 102639.

<https://doi.org/10.1016/j.psychsport.2024.102639>

Zur Veröffentlichung akzeptiert am 4. 4. 2024.

Studie III

Dings, A., & Kubillus, L. K. (2025). Beyond case vignettes: do diagnostic labels affect how symptoms of mental illness are perceived? *Journal of Mental Health Disorders*, 5(1), 1-14.

<https://doi.org/10.33696/mentalhealth.5.030>

Ursprünglich zur Veröffentlichung akzeptiert am 26. 12. 2024, korrigiertes Manuskript akzeptiert am 6. 8. 2025, aktuell *in press*.

1 Einleitung

Unter dem breiten Begriff der (subjektiven oder objektiven) Lebensqualität wird eine Reihe heterogener Komponenten zusammengefasst, die zu einem angenehmen (hedonistischen) und sinnerfüllten (eudämonischen) Leben beitragen (Diener, 1984; Ryan & Deci, 2001; Skevington & Böhnke, 2018). So teilt beispielsweise die Weltgesundheitsorganisation die Lebensqualität in einen globalen Faktor und sechs domänenspezifische Maße auf, die von körperlicher und mentaler Gesundheit über materielle Versorgung bis hin zu gesellschaftlicher Teilhabe reichen (WHOQoL SRPB Group, 2006). Dabei können objektive Maße wie das finanzielle Einkommen oder subjektive Indikatoren, d.h. selbstberichtete Bewertungen der eigenen Lebensqualität, herangezogen werden. Beide Modalitäten korrelieren oft nur moderat miteinander (Cummins, 2000), und in der psychologischen Forschung steht zunehmend die subjektive Lebensqualität im Vordergrund (Camfield & Skevington, 2008). In der vorliegenden Abhandlung werden in methodisch und inhaltlich heterogenen Studien drei Forschungsfragen behandelt, die einzelne Aspekte der Lebensqualität in verschiedenen Kontexten beleuchten.

Die Studien I (Dings & Spinath, 2021) und II (Dings & Spinath, 2024) nutzten dazu Daten des *TwinLife*-Panels, das deshalb im Folgenden kurz vorgestellt wird.

TwinLife

TwinLife wurde als längsschnittliche Panelstudie mit dem Ziel konzipiert, Zwillingspaare vom Kindes- bis zum jungen Erwachsenenalter hinweg in ihrem Lebensweg zu begleiten. Ein Fokus liegt dabei auf Variablen und Entwicklungen, in denen diese Wege divergieren können – unterschiedlicher akademischer und finanzieller Erfolg, Lebensereignisse und ihre Valenz, Sozialleben u. Ä. (Hahn et al., 2016). Damit deckt der TwinLife-Datensatz ähnliche Bereiche ab wie die erwähnten Aspekte der Lebensqualität.

Seit dem Jahr 2014 wurden im Rahmen von TwinLife eineiige sowie gleichgeschlechtliche zweieiige Zwillingspaare in Deutschland begleitet. Dabei wurden neben den Zwillingen selbst auch ihre Eltern, Geschwister und gegebenenfalls weitere Bezugspersonen befragt (extended twin family design; Keller et al., 2010). Die hauptsächlichen Befragungswellen wurden dabei vor Ort, d.h. „face to face“, durch geschulte Interviewer durchgeführt. Zwischen diesen Befragungen wurden weitere Informationen telefonisch oder online erfasst.

Zum Zeitpunkt des ersten Face-to-Face-Interviews nahmen 4097 Familien an TwinLife teil. Diese

Stichprobe wurde erstmals bei einer deutschen Zwillingsstudie durch ein zufälliges Sampling von Familien mithilfe der Einwohnermeldeämter aufgebaut (Mönkediek et al., 2019). Dadurch konnten Selbstselektionseffekte vermieden und eine gute Abdeckung insbesondere verschiedener sozioökonomischer Schichten erreicht werden, wenngleich die TwinLife-Stichprobe nicht perfekt bevölkerungsrepräsentativ ist (so sind Familien mit Migrationshintergrund etwas unterrepräsentiert). Die Zwillingsstichprobe in TwinLife ist in vier Alterskohorten aufgeteilt, die zum Zeitpunkt der ersten Befragungswelle 5, 11, 17 und 23 / 24 Jahre alt waren. Für eine genauere Übersicht über die Merkmale der Stichprobe sei auf Lang und Kottwitz (2020) verwiesen. Die TwinLife-Daten können zum Stand der vorliegenden Abhandlung unter dem Link https://search.gesis.org/research_data/ZA6701 als *scientific use file* abgerufen werden.

2 Schulisches Underachievement

Bei den jüngeren Zwillingskohorten in TwinLife liegt ein inhaltlicher Fokus auf ihrer schulischen Laufbahn. Schulerfolg ist nicht nur während des Heranwachsens selbst ein relevanter Faktor für das eigene subjektive Wohlbefinden (Steinmayr et al., 2016), sondern auch ein Vorläufer einer höheren Lebensqualität im Erwachsenenalter (Edgerton et al., 2011). Aufgrund dieser weitreichenden Konsequenzen liegt eine wichtige Forschungsfrage darin, Fälle sogenannten Underachievements zu erklären: Schüler, die trotz guter allgemeiner Intelligenz keine erfolgreiche Schullaufbahn erreichen. Studie I nutzte die in TwinLife vorhandenen schulischen Variablen, um Charakteristiken solcher Fälle zu identifizieren.

Definition und Operationalisierung

Schulischer Erfolg im Sinne guter Zensuren ist auf eine große Bandbreite an Personen- und Umweltvariablen zurückzuführen (Schneider & Preckel, 2017; Hattie & Yates, 2013). Dabei ist die allgemeine Intelligenz ein naheliegender Indikator dafür, zu welchen Schulleistungen ein Kind oder ein Jugendlicher in der Lage wird, und hat sich in der Tat als bester bekannter einzelner Prädiktor des schulischen und auch allgemeineren akademischen Erfolges erwiesen. Mit einer durchschnittlichen Varianzaufklärung von rund 25% im Schulkontext (Roth et al., 2015) ist es dennoch kein seltener Fall, dass die Noten einzelner Schüler deutlich unter oder deutlich über dem aufgrund ihrer kognitiven Fähigkeiten „erwartbaren“ Niveau liegen. Eine deutliche solche Diskrepanz in negativer Richtung wird als schulisches Underachievement bezeichnet (Reis & McCoach, 2000).

Dabei können aus der inhaltlichen Definition von Underachievement jedoch verschiedene Schlussfolgerungen hinsichtlich seiner statistischen Operationalisierung gezogen werden (Hoover-Schultz, 2005; Ziegler et al., 2012; White et al., 2018; Mazrekaj et al., 2022). Diese führen zu deutlich verschiedenen Fallselektionen (Lau & Chan, 2001; Jackson & Jung, 2022), sodass die Wahl der Methode nicht unerheblich ist.

Die sogenannte *Differenzmethode* vergleicht z-standardisierte Werte in einem Maß für Schulerfolg und einem Maß für kognitive Fähigkeiten innerhalb jeder Person und stellt Underachievement bei einer Diskrepanz (in Richtung schlechterer Schulnoten) von typischerweise 1.0 oder mehr Standardabweichungen fest. Wegen des Risikos dieser Methode, unreliable Fallselektionen aufgrund von Regression zur Mitte mit sich zu bringen (Cone & Wilson, 1981), wird häufiger als die Differenzmethode die *Regressionsmethode* gewählt. Statt einer absoluten Differenz werden dabei die Residuen einer Regression des Maßes für Schulerfolg auf das Maß für kognitive

Fähigkeiten betrachtet und eine Diskrepanz von beispielsweise einem Standardfehler als Schwellenwert für Underachievement genutzt.

Diesen beiden genannten Ansätzen ist jedoch gemeinsam, dass jeder Fall einer Diskrepanz, die den (arbiträren) Cutoff-Wert überschreitet, als Underachievement eingestuft wird. Dies kann dazu führen, dass sehr begabte Schüler auch bei guten Noten als Underachiever klassifiziert werden, da ihnen aufgrund ihrer Intelligenz *sehr* gute Noten prognostiziert worden wären. Es kann als zweifelhaft angesehen werden, ob diese Fälle mit beispielsweise tatsächlich versetzungsgefährdeten Schülern unter einem Oberbegriff zusammengefasst werden sollten (Hanses & Rost, 1998). Um diese Problematik zu vermeiden, kann Underachievement alternativ nicht aufgrund einer bestimmten relativen Diskrepanz, sondern als Konstellation aus zwei kategorisch abgegrenzten Merkmalen definiert werden: überdurchschnittliche kognitive Fähigkeiten bei gleichzeitig nicht nur *relativ* schwächeren, sondern tatsächlich unterdurchschnittlichen Schulleistungen. Diese Vorgehensweise wird auch als *absolute split method* bezeichnet (Lau & Chan, 2001). Die erhoffte validere Fallselektion durch diese Methode „erkauft“ man sich dabei durch einen Verlust an Varianz aufgrund der Dichotomisierung der Daten (Kawakami et al., 2024).

Studie I wurde mit dem Ziel durchgeführt, diese *absolute split method* weitergehend umzusetzen als in der bisherigen Forschung überwiegend der Fall. Auch diejenigen Studien, die die *absolute split method* nutzen, vergleichen diese Gruppe an Schülern in der Regel ausschließlich mit ähnlich kognitiv leistungsfähigen, aber auch tatsächlich erfolgreichen Schülern. Dieser Vergleich ist jedoch durch den Schulerfolg konfundiert (McCall et al., 2000): Unterschiede zwischen diesen beiden Gruppe könnten Korrelate des Schulerfolgs im Allgemeinen kennzeichnen, statt spezifisch für Underachievement als *erwartungswidrigem* Misserfolg zu sein. Um diese Konfundierung zu vermeiden, ist ein gleichzeitiger Vergleich von Underachievern mit anderen ebenfalls nicht erfolgreichen Schülern wünschenswert, wie dies in Tabelle 1 illustriert ist.

Tabelle 1. Einteilung der Schüler nach der *absolute split method* in Studie I

	Überdurchschnittliche Intelligenz	Mittlere Intelligenz	Unterdurchschnittliche Intelligenz
Überdurchschnittliche Schulleistungen	erwartungsgemäße Hochleister	<i>nicht betrachtet</i>	Overachiever
Durchschnittliche Schulleistungen	<i>nicht betrachtet</i>	<i>nicht betrachtet</i>	<i>nicht betrachtet</i>
Unterdurchschnittliche Schulleistungen	Underachiever	<i>nicht betrachtet</i>	erwartungsgemäße Niedrigleister

Ein solches Schema ist trotz frühen dahingehenden Empfehlungen in der Literatur (McCall & Lau, 2000; McCoach & Siegle, 2003a) bisher selten umgesetzt worden oder nur ansatzweise (siehe McCall et al., 2000; Hanses & Rost, 1996; Vlahovic-Stetic et al., 1999). Zudem bezieht das Schema in Tabelle 1 auch das spiegelbildliche Phänomen Overachievement ein, das bisher vergleichsweise wenig Aufmerksamkeit in der Forschung erhalten hat (Sparfeldt et al., 2010).

Wie auch bei der Regressions- oder der Differenzmethode bleibt die Frage nach sinnvollen Cutoff-Werten bei der Identifikation von Underachievement per *absolute split method* im Prinzip arbiträr (Ziegler et al., 2012). Folgt man jedoch dem Ansatz vieler Studien mit der Differenzmethode, nach dem die z-standardisierten Werte der beiden Maße um mindestens eine Standardabweichung voneinander abweichen sollen, lässt sich dies in der *absolute split method* so umsetzen, dass die kognitiven Fähigkeiten um mindestens 0.5 Standardabweichungen über und die Schulleistungen um mindestens 0.5 Standardabweichungen unter dem Durchschnitt liegen sollten. In Studie I wurde zudem die angestrebte inhaltliche Validität versucht zu erreichen, indem nur diejenigen Fälle in TwinLife als Underachiever identifiziert wurden, die sich in mindestens einem der beiden herangezogenen Fächer (Mathe und Deutsch) mit einer Note von 4 oder schlechter nahe an einer Versetzungsgefährdung bewegten. Das Kriterium „mindestens 0.5 Standardabweichungen unter dem Durchschnitt“ erfüllten in der Stichprobe teilweise auch Schüler mit der Note 3 in beiden Fächern, da Schulnoten in Deutschland sehr stark in Richtung guter Noten schiefverteilt sind (siehe Schleithoff, 2015).

Insgesamt lagen von 612 Gymnasiasten (324 weiblich, 288 männlich) mit einem mittleren Alter von 14.6 Jahren ($SD = 2.8$) Angaben in allen relevanten Variablen in der ersten Erhebungswelle von TwinLife vor. Auf Gymnasiasten war die Untersuchung beschränkt, da ein sinnvoller Vergleich des Erfolgskriteriums – der Schulnoten – über die verschiedenen weiterführenden Schulformen im deutschen Schulsystem hinweg nicht möglich ist (siehe Uhlig et al., 2009).

Von diesen 612 Schülern erfüllten 385 (63%) keine der Kriterien, die sie einer der definierten Gruppen von Underachievement, erwartungsgemäße Niedrigleister, erwartungsgemäße Hochleister oder Overachievement zugeordnet hätten, wiesen also entweder bei den kognitiven Fähigkeiten oder bei der Schulleistung einen Wert auf, der sie im Mittelfeld (zwischen -0.5 und +0.5 Standardabweichungen) einer oder beider dieser Variablen platzierte. Unter den verbliebenen 227 Personen wurden nach Anwendung der oben beschriebenen Kriterien 21 Schüler und somit rund 3% der Gesamtstichprobe der Underachievement-Gruppe zugeordnet. Von den als überdurchschnittlich (mindestens 0.5 Standardabweichungen über dem Durchschnitt liegenden) intelligenten Schülern wurden rund 10% als Underachiever klassifiziert. Beide Werte zeigen eine im Vergleich zu anderen Studien recht strenge Fallselektion an (siehe White et al., 2018). Dies

verringert gegenüber liberaleren Auswahlkriterien den Anteil an nicht validen Fällen in der Stichprobe, die aufgrund der Unreliabilität von Prädiktor und Kriterium als Artefakte zu erwarten sind (Ziegler et al., 2012). Ähnliche Zahlen ergeben sich für Overachievement: dieser Gruppe wurden 30 Personen zugeordnet. Die Gruppen der erwartungsgemäßen Niedrig- und Hochleister machten 63 beziehungsweise 113 Personen aus.

Die 227 Schüler in den Subgruppen kamen größtenteils aus unterschiedlichen Familien, jedoch waren auch 17 eineiige und 20 zweieiige Zwillingspaare vorhanden. Die 21 Underachiever in der Stichprobe stammten aus 21 verschiedenen Familien.

Das Achievement Orientation Model (AOM)

Unter verschiedenen theoretischen Modellen, welche die Risikofaktoren und Begleiterscheinungen von Underachievement zusammenzufassen versuchen (siehe Siegle, 2018; Figg et al., 2012; White et al., 2018), hat unter anderem das *Achievement Orientation Model*, kurz AOM (Siegle & McCoach, 2005) gute empirische Bestätigung erfahren. Es baut auf allgemeinen Motivationstheorien auf (siehe Rubenstein et al., 2012) und sieht motivationale Variablen als notwendige Voraussetzung dafür, dass gegebene kognitive Fähigkeiten auch zu Schulerfolg führen können. Im Kern stehen hierbei die Konstrukte *goal valuation* (Zielvalenz), *self-efficacy* (Selbstwirksamkeit) und *environmental perceptions* (Wahrnehmung des sozialen Umfeldes). Mit anderen Worten müssen Schüler den schulischen Erfolg erstens überhaupt für ein lohnenswertes Ziel halten, müssen zweitens Vertrauen in die eigenen Fähigkeiten haben, dieses Ziel zu erreichen und sich drittens bei dem Versuch dazu vom Umfeld, etwa Eltern und Lehrkräften, unterstützt fühlen. Sind diese Voraussetzungen gegeben, postuliert das AOM zudem, dass Schüler zudem über erfolgreiche selbstregulative Strategien in der schulischen Praxis verfügen müssen. Dazu zählen realistische Zielsetzungen und effektive Lernstrategien (Siegle & McCoach, 2005).

Zu den empirischen Belegen für die Validität des AOMs zählen insbesondere Befunde, dass (hochbegabte) Underachiever tatsächlich geringere Werte in diesen motivationalen Komponenten berichten als (hochbegabte) erfolgreiche Schüler (Vlahovic-Stetic et al., 1999; McCoach & Siegle, 2003b; Balduf, 2009; Landis & Reschly, 2013; Castejon et al., 2016; White et al., 2018; Mofield & Parker Peters, 2019). Auch die medierende Rolle der selbstregulativen Fähigkeiten, d. h. beispielsweise guter Lernstrategien, hat sich wiederholt bestätigt (Ritchotte et al., 2014; Rubenstein, et al., 2012; Veas Iniesta et al., 2017). Allerdings beschränken sich diese Untersuchungen auf den Vergleich innerhalb hochbegabter Schüler, sodass der „Quervergleich“ zu anderen nicht erfolgreichen, aber gleichzeitig auch nicht hochbegabten Schülern fehlt (siehe Tabelle 1). Studie I

wurde mit dem Ziel durchgeführt, die Konstrukte aus dem AOM in einer solchen umfassenderen Betrachtung zu überprüfen.

Das AOM kann mit den vorliegenden Paneldaten nicht perfekt operationalisiert werden, da in TwinLife nicht auf dieses konkrete Modell Bezug genommen wird. Der für das AOM entwickelte spezifische Fragebogen, das *School Attitude Assessment Survey* (McCoach & Siegle, 2003b), ist somit nicht enthalten. Jedoch sind einige motivationale Konstrukte, die in TwinLife erfasst wurden, den AOM-Komponenten inhaltlich sehr ähnlich.

So ist das Konstrukt Selbstwirksamkeit in Form der *Allgemeinen Selbstwirksamkeit Kurzskala* (ASKU; Beierlein et al., 2012) unmittelbar in TwinLife enthalten. Zudem wurde als spezifischer schulbezogenes Konstrukt auch das häufig mit Underachievement korrelierende akademische Selbstkonzept berücksichtigt (Skala nach Dickhäuser et al., 2012). Auch die AOM-Komponente *goal valuation* ist mit den *Skalen zur Erfassung der Lern- und Leistungsmotivation* (SELLMO; Spinath et al., 2002) gut abgebildet.

Zur Annäherung an das Konstrukt *environmental perceptions* wurde in Studie I der kindberichtete Erziehungsstil der Eltern genutzt, zu dem elterliche Unterstützung und Autonomie als positive sowie elterliche Kontrolle als negativer Faktor zählten. Damit ist ein Teilaspekt der Unterstützung durch die Umwelt des Kindes abgedeckt, wenngleich die Unterstützung beispielsweise durch Lehrpersonen nicht erfasst ist. Jeder der drei Aspekte des Erziehungsstils wurde mit je drei Items nach Spinath und Wolf (2006) erfasst.

Als Proxy für die selbstregulativen Fähigkeiten, die im AOM als Mediator zwischen den motivationalen Variablen und Schulerfolg fungieren, wurde die Subskala „Hyperaktivität“ einer längeren Skala zu externalisierendem Verhalten genutzt (adaptiert nach Goodman et al., 1998). Auch diese Variable kann nur als unvollständige Erfassung des entsprechenden AOM-Konstruktes angesehen werden. Frühere Untersuchungen haben gesteigerte Hyperaktivität jedoch bereits als Bestandteil der fehlenden selbstregulativen Eigenschaften bei Underachievern festgestellt (McCoach & Siegle, 2020; Barry et al., 2002).

Neben den Konstrukten aus dem AOM wurden in Studie I zudem die Persönlichkeitsfaktoren des Big-Five-Modells (Costa & McCrae, 1992) berücksichtigt, da zu diesen sehr prominenten Traits vergleichsweise wenige Erkenntnisse im Kontext von Underachievement vorliegen (Reis & McCoach, 2000; Preckel et al., 2006). Die Big-Five-Faktoren liegen in TwinLife in Form des *Big Five Inventory-SOEP* vor (BFI-S; Gerlitz & Schupp, 2005).

Die kognitiven Fähigkeiten wurden in TwinLife mithilfe des *Culture Fair Tests* (CFT) gemessen. Kindern bis 9 Jahren wurden die Untertests „Figurales Denken“, „Figurale Wahrnehmung“ und „Matrizen“ aus dem CFT 1-R (Weiß & Osterland, 2012) vorgelegt, Versuchspersonen ab 10 Jahren

alle vier Untertests der weiterentwickelten Form CFT 20-R (Weiß & Weiß, 2006). Inhaltlich handelt es sich bei den Aufgaben um möglichst stark um kulturell bedingte Einflüsse bereinigte Tests zur Messung der fluiden Intelligenz nach Cattell, insbesondere schlussfolgernden Denkens. Details zur Intelligenzmessung in TwinLife sowie der weiteren Aufbereitung der Testdaten sind bei Gottschling (2017) dargestellt.

Es wurden die jeweils jüngsten Zeugnisnoten der Schüler in Mathematik und Deutsch als Kriterien der Schulleistung herangezogen. Aufgrund der Tatsache, dass (nur) diese Fächer in praktisch allen Schulen gelehrt werden, ist die Eingrenzung auf diese beiden Fächer nicht nur ein gängiger Ansatz in der Schulleistungsforschung im deutschsprachigen Raum (siehe etwa Stanat et al., 2017), sondern auch ein praktisch notwendiger, da alle anderen Fächer in der TwinLife-Stichprobe nur teilweise vorhanden waren.

Ergebnisse und Schlussfolgerungen

Zur Untersuchung, welche Variablen die vier Gruppen aus Tabelle 1 spezifisch kennzeichneten, wurde eine Diskriminanzanalyse durchgeführt. Die Daten verletzen die dafür notwendige Annahme homogener Kovarianzen nicht (Box's M = 286.8, $p = 0.357$).

Die ersten beiden der drei möglichen Diskriminanzfunktionen trugen signifikant zur Varianzaufklärung bei, wobei das kanonische R^2 der ersten Funktion bei 0.26, das der zweiten Funktion bei 0.14 lag.

In der ersten Diskriminanzfunktion lagen die Gruppenzentroide der Underachiever (-0.79) und der erwarteten Niedrigleister (-0.75) nah beieinander. Ihnen standen Gruppenzentroide von 0.45 bei den erwarteten Hochleistern und 0.44 bei den Overachievern gegenüber. Die erste Funktion trennte somit ganz überwiegend die beiden schulisch erfolgreichen von den weniger erfolgreichen Gruppen, teilte die Stichprobe also entlang des Leistungskriteriums auf und unterschied dabei nur minimal zwischen „erwartetem“ und „unerwartetem“ Erfolg oder Misserfolg.

Inhaltlich zeigen die kanonischen Korrelationen der einzelnen Variablen hier plausible Zusammenhänge, welche den bisherigen Literaturbefunden zu Prädiktoren von Schulerfolg entsprechen: ein höheres akademisches Selbstkonzept (kanonische Korrelation: 0.59), höhere Selbstwirksamkeit (0.48), Lernmotivation (0.40) und Gewissenhaftigkeit (0.58) sowie geringere Hyperaktivität (-0.46) kennzeichnen die beiden schulisch erfolgreichen Gruppen im Vergleich zu den weniger erfolgreichen. Auch ein konstruktiverer Erziehungsstil mit hoher Autonomie (0.23) und Unterstützung (0.26), aber weniger harscher Kontrolle (-0.51) unter den erfolgreichen Schülern entspricht den Erwartungen aufgrund bisheriger Forschungen und den Annahmen des AOM.

Über diese Aufteilung entlang des Leistungskriteriums hinaus kann die zweite Funktion der Diskriminanzanalyse zusätzlich signifikant Varianz in der Gruppenzugehörigkeit aufklären. Wie die Gruppenzentroide dieser Funktion zeigen, erfolgt diese zusätzliche Varianzaufklärung entlang der Schulleistung *in Relation* zu den kognitiven Fähigkeiten: Underachievement (-0.73) und Overachievement (0.77) bilden die beiden Extrema dieser Funktion, während erwartungsgemäße Niedrigleistung (0.25) und erwartungsgemäße Hochleistung (-0.21) sich dazwischen gruppieren. Der signifikante Beitrag dieser Funktion lässt sich somit so auffassen, dass hier Faktoren identifiziert werden, die tatsächlich spezifisch mit Underachievement als *relativem* schulischem Misserfolg assoziiert sind sowie spiegelbildlich dazu mit Overachievement.

Die kanonischen Korrelationen zeigen hier die höchsten Zusammenhänge nicht mit den aus dem AOM übernommenen Variablen an, sondern mit einigen der Big Five: Underachievement geht mit niedrigerer Extraversion einher (0.33; man beachte, dass dieser positive Wert angesichts der Richtung der Gruppenzentroide eine geringere Extraversion in der Gruppe der Underachiever bedeutet) sowie mit niedrigerer Gewissenhaftigkeit (0.31) und mit *höherer* Offenheit für Erfahrungen (-0.30). Gering bis moderat ausgeprägte kanonische Korrelationen legen außerdem eine geringere Verträglichkeit (0.26) und Lernmotivation (0.22) in der Gruppe der Underachiever nahe. Dagegen ist das akademische Selbstkonzept nur noch in geringem Ausmaß (0.12) mit dieser Diskriminanzfunktion korreliert. Auch die Variablen des Erziehungsstils sind mit kanonischen Korrelationen zwischen -0.04 (Kontrolle) und 0.18 (Autonomie) deutlich weniger relevant als in der 1. Diskriminanzfunktion.

Mithilfe der Diskriminanzfunktionen konnten 36.1% der Fälle korrekt klassifiziert werden, wenn *leave-one-out*-Klassifikationen genutzt wurden, jeder Fall also ausschließlich aufgrund der restlichen Stichprobe vorhergesagt wurde (Molinaro et al., 2005). Mit regulärer statt *leave-one-out*-Klassifikation lag dieser Wert deutlich höher (45.8%). Dies deutet auf starkes Overfitting der Diskriminanzfunktionen hin, was angesichts der Kombination aus kleinen Gruppengrößen und relativ dazu einer hohen Zahl an Prädiktoren als plausibel einzuordnen ist. Die Ergebnisse sind somit als wenig robust zu betrachten, solange sie nicht in anderen Stichproben repliziert werden. Eine Wiederholung der Analyse innerhalb der Gesamtschüler statt Gymnasiasten in der TwinLife-Stichprobe zeigte jedoch insgesamt ähnliche Ergebnisse.²

Bei diesen Ergebnissen ist zu beachten, dass sich die Interpretation der Diskriminanzanalyse nicht nur auf die Gruppe der Underachiever und ihren Vergleich mit erwarteter Hoch- und Niedrigleistung stützt, sondern ebenso auf die Kontrastierung dieser beiden letztgenannten Gruppen mit der Gruppe der Overachiever. Die beiden Gruppen der Under- und Overachiever erscheinen in

² Diese zusätzliche Analyse war nicht Teil der als Studie I publizierten Arbeit.

den Ergebnissen somit als zwei Seiten einer Medaille: die Gruppenzentroide der Overachiever-Subgruppe in den beiden Diskriminanzfunktionen verhalten sich zu den beiden „mittleren“ Gruppen der erwartungsgemäßen Niedrig- und Hochleister nahezu exakt spiegelbildlich zu jenen der Underachiever-Subgruppe. Auch in den deskriptiven Daten der Variablen, die die Gruppenzugehörigkeit besonders deutlich trennen, deutet sich dieses symmetrische Bild an (Tabelle 2): beispielsweise zeigen nicht nur Underachiever besonders niedrige Werte in Gewissenhaftigkeit, sondern die Gruppe der Overachiever weist hier umgekehrt besonders hohe Werte gegenüber den beiden weiteren Vergleichsgruppen auf. Dies ist nicht bei allen betrachteten Variablen der Fall, insbesondere die Diskriminanzfunktionen lassen jedoch vermuten, dass nicht nur Underachievement, sondern ebenso auch Overachievement mithilfe der motivationalen und persönlichkeitsbezogenen Konstrukte deutlich von den beiden weiteren Vergleichsgruppen getrennt werden kann.

Tabelle 2. Mittelwerte und Standardabweichungen der Prädiktorvariablen in den für die Studie definierten Gruppen

	Underachiever (n = 21)	Erwartungsgemäße Niedrigleister (n = 63)	Erwartungsgemäße Hochleister (n = 113)	Overachiever (n = 30)
Erziehungsstil:	9.24 (2.12)	8.86 (2.74)	9.75 (2.27)	9.61 (2.71)
Unterstützung Erziehungsstil:	8.71 (2.63)	9.03 (2.61)	9.53 (2.41)	10.05 (2.37)
Autonomie Erziehungsstil:	4.71 (2.12)	4.79 (3.22)	2.99 (3.03)	2.73 (2.75)
Kontrolle Lernmotivation	8.81 (1.83)	8.84 (1.88)	9.54 (1.61)	10.14 (1.63)
Selbstwirksamkeit	7.86 (1.93)	8.37 (2.08)	9.18 (1.50)	9.47 (1.76)
Akademisches Selbstkonzept	6.58 (0.67)	6.83 (1.33)	7.52 (1.06)	7.57 (0.77)
Hyperaktivität	3.71 (2.05)	3.88 (2.41)	2.58 (1.93)	3.05 (2.18)
Neurotizismus	11.86 (4.33)	11.71 (5.11)	11.17 (4.68)	11.17 (4.01)
Extraversion	10.57 (5.04)	12.22 (4.20)	11.27 (3.67)	12.27 (3.62)
Gewissenhaftigkeit	10.00 (2.92)	11.16 (2.96)	12.74 (2.82)	13.43 (3.17)
Verträglichkeit	13.00 (2.05)	13.56 (3.44)	13.69 (2.89)	14.57 (2.53)
Offenheit	17.19 (4.26)	16.06 (4.06)	17.65 (3.97)	16.43 (3.97)

Anmerkungen. Wertebereiche der Skalen: 0-12 für die Erziehungsstil-Skalen sowie Lernmotivation, Selbstwirksamkeit und akademisches Selbstkonzept; 0-10 für Hyperaktivität; 0-18 für Extraversion, Gewissenhaftigkeit und Verträglichkeit sowie 0-24 für Neurotizismus und Offenheit.

Anhand der Distanz zwischen den Gruppenzentroiden kann untersucht werden, welcher der beiden Vergleichsgruppen die Gruppe der Underachiever in den untersuchten Variablen ähnlicher ist. Diese Auswertung verortet die Underachiever etwas näher an der Gruppe der erwartungsgemäßen Niedrigleister, denen sie im Hinblick auf den fehlenden Schulerfolg ähneln (Distanz der

Gruppenzentroide = 0.97) als an der Gruppe der erwartungsgemäßen Hochleister, denen sie im Hinblick auf die kognitiven Fähigkeiten ähneln (1.34). Zudem kann im Rahmen der Gruppenklassifikation betrachtet werden, welche Wahrscheinlichkeit jeder Person zugeschrieben wird, zu den vier Subgruppen zu gehören. Tatsächlich Underachievern wurde hier mit einer etwas höheren Wahrscheinlichkeit (26%) eine Zugehörigkeit zu der Gruppe der erwartbaren Niedrigleister zugeschrieben als zu der Gruppe der erwartbaren Hochleister (20%). Dieser Unterschied war mit 25% gegenüber 23% nur noch minimal, wenn das akademische Selbstkonzept nicht berücksichtigt wurde (da argumentiert werden könnte, dass diese Angabe bereits teilweise „verrät“, ob ein Schüler erfolgreich ist oder nicht, wenngleich akademisches Selbstkonzept und Schulerfolg in der vorliegenden Stichprobe nur moderat mit $r = 0.36$ korrelierten). Auch dieses Ergebnis zeigte sich spiegelbildlich in der Gruppe der Overachiever, die mit 26% gegenüber 20% (beziehungsweise 25% gegenüber 21% ohne Berücksichtigung des akademischen Selbstkonzepts) etwas eher der Bezugsgruppe mit ähnlich guter schulischer Leistung als der Bezugsgruppe mit ähnlich unterdurchschnittlichen kognitiven Fähigkeiten zugeordnet wurden. Insgesamt deuten diese Ergebnisse aber darauf hin, dass die Gruppen der Over- und Underachiever keiner der beiden Bezugsgruppen deutlich ähnlicher sind als der anderen. Dies entspricht den Ergebnissen aus einer der wenigen bestehenden ähnlichen Studien (Hanses & Rost, 1996).

Studie I testete zudem zwei weitere theoretische Annahmen über Underachievement. In Tabelle 2 ist deskriptiv erkennbar, dass die Gruppe der Underachiever in den meisten untersuchten Variablen keine höhere Varianz aufweisen als die übrigen Gruppen. Levene-Tests zu den einzelnen Variablen zeigten signifikant verschiedene Varianzen über die vier Gruppen hinweg bei Selbstwirksamkeit, akademischem Selbstkonzept, Verträglichkeit und Extraversion, von denen jedoch nur Extraversion die numerisch größte Varianz in der Gruppe der Underachiever aufwies. Dies wurde betrachtet, um eine verschiedentlich in der Literatur zu findende Annahme über Underachiever zu überprüfen: „As a group, underachievers differ more from each other than achievers differ from each other“ (Siegle, 2018, S. 285). Anders als einzelne frühere Untersuchungen (McCoach & Siegle, 2003a; Smith, 2007) bestätigen die deskriptiven Ergebnisse aus Studie I dies somit nicht, zumal auch die zusammenfassenden Scores aus der Diskriminanzanalyse keine signifikant verschiedene Varianz je nach Subgruppe aufwiesen.

Studie I testete zudem die Annahme, dass die motivationalen Faktoren im AOM nicht untereinander kompensiert werden können. Ritchotte und Kollegen (2014) führen als Beleg für diese Annahme an, dass in ihren Daten Underachiever deutlich wahrscheinlicher als die Vergleichsgruppe geringe Werte in mindestens einer der Komponenten aufwies. Einen direkten Test dafür, ob hohe Werte in einer der Komponenten niedrige Werte in einer anderen Komponente ausgleichen können, stellt

dies jedoch nicht dar. Ist eine Kompensation nicht möglich, sollten das „Fehlen“ einer der Komponenten inkrementelle Validität in der Vorhersage von Underachievement über den Mittelwert der drei Komponenten hinaus aufweisen.

Um eine solche, die Interaktion zwischen Komponenten betreffende mit möglichst großer statistischer Power testen zu können, wurde anstelle des restriktiven Selektionsansatzes an dieser Stelle auf die „kontinuierliche“ Operationalisierung von Underachievement als Residuum einer Regression von Schulnoten auf kognitive Testwerte zurückgegriffen. Höhere Werte stehen hier für höheres Underachievement, d.h. eine niedrige Schulleistung im Vergleich zum vorhergesagten Wert. In Anlehnung an Ritchotte et al. (2014) wurden Werte, die 1.0 Standardabweichungen oder mehr unter dem Mittelwert des jeweiligen Konstruktes lagen, als „Fehlen“ einer motivationalen Komponente definiert. Zudem wurde der geringste Wert in den drei Komponenten bei der jeweiligen Person als Maß dafür herangezogen, ob mindestens einer der motivationalen Faktoren gering ausgeprägt ist.

Der Mittelwert der drei z-standardisierten Komponenten korrelierte deutlich mit dem regressionsbasierten Underachievement-Wert ($r = 0.37$, $p < 0.001$). Auch das intraindividuelle Minimum der Motivationskomponenten sowie die Anzahl der „fehlenden“ Komponenten zeigten bivariat erwartungsgemäße Korrelationen ($r = 0.27$ bzw. $r = -0.37$; p jeweils < 0.001). Bei Gelten der Kompensationsannahme sollten diese Maße über den Mittelwert der Komponenten hinaus inkrementellen Vorhersagewert beibehalten. Dies war in einer gemeinsamen Regression jedoch nicht der Fall ($r = -0.09$, $p = 0.17$ für das Minimum der Komponenten und $r = -0.11$, $p = 0.08$ für die Anzahl der gering ausgeprägten Komponenten). Mit den vorliegenden Daten finden sich somit keine Hinweise darauf, dass die spezifische Annahme einer nicht möglichen Kompensation zwischen den AOM-Faktoren erfüllt ist. Bei dieser Schlussfolgerung muss jedoch bedacht werden, dass manche der Variablen in Studie I inhaltlich nur näherungsweise den AOM-Konstrukten entsprachen.

Die Ergebnisse von Studie I lassen gemischte Schlussfolgerungen zum Konstrukt Underachievement zu. Zunächst können die Ergebnisse als Bestätigung zahlreicher früherer Erkenntnisse gelten, nach denen Schulerfolg typischerweise mit einem positiven Blick auf die eigenen Fähigkeiten, auf die Wertigkeit von Schulerfolg als persönliches Ziel sowie mit Unterstützung durch die Eltern in Form eines konstruktiven Erziehungsstils einhergeht. Darüber hinaus sollte in Studie I jedoch auch untersucht werden, ob das Fehlen dieser motivationalen Voraussetzungen generell mit schulischem Misserfolg assoziiert ist oder insbesondere für die Konstellation des Underachievements - Misserfolg trotz überdurchschnittlicher kognitiver Fähigkeiten - typisch ist. Dies bleibt in den Ergebnissen von Studie I deutlich

zweifelhafter. Nachdem in der 1. Diskriminanzfunktion überwiegend die Unterschiede zwischen niedrig- und hochleistenden Schülern im Allgemeinen abgebildet wurden, ergaben sich in der 2. Diskriminanzfunktion nur noch deutlich geringere kanonische Korrelation für Lernmotivation und akademisches Selbstkonzept. Dagegen wurden hier einige der Big-Five-Faktoren relevanter: Underachievement in Abgrenzung zu „erwartbarem“ schulischem Misserfolg ging tendenziell mit niedrigeren Werten in Verträglichkeit, Gewissenhaftigkeit und Extraversion einher sowie mit höherer Offenheit für Erfahrungen. Aufgrund der geringen Fallzahl insbesondere in der Gruppe der Underachiever ist dieses neuartige Ergebnis mit Vorsicht zu interpretieren, zumal sich bei Anwendung einer Leave-one-out-Klassifikation deutliche Hinweise auf mögliches Overfitting fanden. Gerade der Befund, dass geringe Extraversion kennzeichnend für Fälle von Underachievement sein könnte, deckt sich jedoch mit weiteren empirischen Hinweisen darauf, dass introvertierte Schüler im Schulkontext oft unter ihren Möglichkeiten bleiben (Godfrey & Koutsouris, 2023; von Stumm & Plomin, 2018).

Die geringen Effekte der AOM-Variablen in der 2. Diskriminanzfunktion lassen es plausibel erscheinen, dass die Unterschiede in diesen Komponenten bei den in der Literatur gängigen Vergleichen zwischen Underachievern und ausschließlich ebenfalls begabten, aber erfolgreichen Schülern zumindest teilweise durch eine Konfundierung dieses Vergleichs mit dem Schulerfolg selbst erklärt werden können. Gerade in Bezug auf akademisches Selbstkonzept ist es nicht verwunderlich, dass Schüler mit schlechten Noten auf dem Zeugnis wenig Vertrauen in ihre eigenen schulischen Fähigkeiten berichten. Zwar ist neben diesem beinahe trivialen Effekt der Schulleistungen auf das schulische Selbstkonzept durchaus auch ein kleinerer Effekt in die andere Richtung längsschnittlich gut nachgewiesen (Wu et al., 2021), dennoch untermauert Studie I, dass es verkürzt wäre, die spezifische Subgruppe der Underachiever auf mangelndes Vertrauen in die eigenen Fähigkeiten oder sonstige Motivationsprobleme zu reduzieren: ein geringes akademisches Selbstkonzept scheint demnach eine Begleiterscheinung aller Formen schulischen Misserfolgs zu sein, nicht nur spezifisch von Fällen, bei denen ein Schüler im Vergleich zu seiner allgemeinen Intelligenz als Underachiever einzustufen ist.

Gleichwohl stützen die Ergebnisse von Studie I allgemeiner gesprochen einen wichtigen Grundbaustein der Forschung zu Underachievement, nämlich die selten explizit diskutierte Annahme, dass die Verdinglichung (Reifikation; siehe Ziegler et al., 2012) von Underachievern als distinkter Subgruppe überhaupt gerechtfertigt ist. Es ist zunächst einmal eine triviale Feststellung, dass ein nicht perfekt reliabel gemessenes Kriterium (Schulleistungen) in vielen Fällen von der Vorhersage durch einen ebenfalls nicht perfekt reliabel gemessenen Prädiktor (kognitive Fähigkeiten) abweichen wird (Ziegler et al., 2012). Underachiever als eigenständige Gruppe mit

bestimmten Eigenschaften zu definieren, ist allein aus der Existenz solcher Fälle somit nicht gerechtfertigt. Vor diesem Hintergrund stützt die Tatsache, dass alle vier Subgruppen in Studie I sich empirisch voneinander abgrenzen ließen, die theoretische Fundierung des Konzepts. Gleichzeitig ist dies nicht der erste Beleg dafür, dass die Klassifizierung bestimmter Fälle als Underachievement nicht nur eine fälschliche Etikettierung von Fehlervarianz darstellt. Neu ist in Studie I der methodische Ansatz, mit dem diese Schlussfolgerung unterstützt wird. Frühere längsschnittliche Studien konnten jedoch bereits zeigen, dass sich Underachievement in vielen Fällen verstetigt und eine Regression zur Mitte von einem Messzeitpunkt zum anderen nicht in dem Ausmaß stattfindet, das zu erwarten wäre, wenn es sich um reine statistische Artefakte handelte (McCall et al., 1992; Colangelo et al., 1993; Johnson et al., 2006).

Zwei weitere in Studie I getestete theoretische Annahmen konnten nicht bestätigt werden: weder zeigte die Subgruppe der Underachiever besonders große Heterogenität im Sinne einer größeren Varianz in den gemessenen Variablen, noch fanden sich Hinweise darauf, dass die motivationalen Komponenten im AOM untereinander nicht kompensiert werden können.

Ziegler und Kollegen (2012) weisen darauf hin, dass eine Abweichung zwischen dem durch kognitive Fähigkeiten vorhergesagten Schulerfolg und den tatsächlichen Noten statistisch nur eines bedeute: dass dieses simple Modell nicht ausreicht, um das Kriterium in diesem Fall korrekt vorherzusagen. In diesem Sinne sollte zukünftige Forschung das Ziel haben, bessere Vorhersagen dafür zu liefern, unter welchen Bedingungen welche Schüler bestmöglich erfolgreich sein können. Dies gilt insbesondere, da Lehrkräfte Underachievement oft nicht als solches erkennen (Rost & Hanses, 1998; Jones & Myhill, 2004; Jackson & Jung, 2022) und häufig fälschlich annehmen, den entsprechenden Schülern fehlte es an Motivation (Siegle et al., 2020). Weitere Forschung könnte ein besseres Verständnis insbesondere der längsschnittlichen Prozesse darüber liefern, wie sich schulischer Misserfolg gerade auch bei guten kognitiven Fähigkeiten entwickelt und verstetigt.

3 Twin-Difference-Analyse der Zusammenhänge zwischen subjektiver Lebenszufriedenheit und dem Besuch von Sportvereinen

Methoden der Zwillingsforschung und das *missing environment*

Während Studie I sich der TwinLife-Daten in erster Linie im Sinne eines Panels bediente, ohne unmittelbar auf die genetisch informierte Natur des Datensatzes zurückzugreifen, machte sich Studie II diesen zentralen Aspekt des TwinLife-Projekts zunutze. An dieser Stelle soll deshalb zunächst allgemeiner auf Methoden und Erkenntnisse der Zwillingsforschung eingegangen werden. Eineiige Zwillingspaare stellen aufgrund ihres (annähernd) identischen Genoms eine Konstellation an Vergleichbarkeit dar, die sich außerhalb dieses „natürlichen Experiments“ schwer finden oder herstellen lässt (McAdams et al., 2021). Dadurch stellen sie eine wichtige Ressource für die differentialpsychologische Forschung dar.

Dabei ermöglicht die gemeinsame Betrachtung mit zweieiigen Zwillingspaaren, die im Durchschnitt zu 50% genetisch übereinstimmen, unter Annahme bestimmter Voraussetzungen (siehe Rijdsdijk & Sham, 2002) eine Zerlegung interindividueller Unterschiede innerhalb der jeweiligen Stichprobe in verschiedene Varianzanteile: (additive) genetische Varianzanteile, sich innerhalb eines Zwillingspaares identisch auswirkende („geteilte“) Umwelteinflüsse und sich ungleich auswirkende („nicht geteilte“) Umwelteinflüsse.³ Dies kann näherungsweise mithilfe von Faustformeln (Falconer, 1965) oder präziser mithilfe von Strukturgleichungsmodellen geschehen (Røysamb & Tambs, 2016). Die Logik dieses „klassischen Zwillingsmodells“ kann auch auf die Kombination mehrerer Variablen angewandt werden, um die Ätiologie ihrer Zusammenhänge zu untersuchen (Posthuma, 2009).

Diese Modelle beantworten jedoch nur die Frage nach dem Ausmaß der genetisch und umweltbedingt erklärbaren Varianz in der jeweiligen Stichprobe, nicht nach inhaltlichen Erklärungen. In jüngerer Zeit konzentriert sich deshalb immer mehr Zwillingsforschung auf eine konkrete Ursachenfindung (siehe Uffelmann et al., 2021; Harden, 2021). Hinsichtlich der genetischen Faktoren zählen dazu der Versuch, bestimmte Outcomes auf bestimmte Allelvarianten (typischerweise *single nucleotide polymorphisms*, kurz SNPs) zurückzuführen, und die Identifikation von bestimmten „Risikogenen“, die in Interaktion mit bestimmten Umweltfaktoren mit spezifischen Outcomes assoziiert sind (Willoughby et al., 2023). Trotz großer Fortschritte in der Machbarkeit und dem Umfang solcher Studien lässt sich festhalten, dass die Entschlüsselung

3 Neben diesem Standardmodell ist auch die Modellierung von dominanten (nicht-additiven) genetischen Einflüssen statt der geteilten Umwelt möglich; beide Komponenten können jedoch alleine mithilfe von eineiigen und zweieiigen Zwillingspaaren nicht gleichzeitig modelliert werden.

solcher Zusammenhänge sich historisch als schwierig erwiesen hat. Wie Harden (2021) in einer Überblicksarbeit darlegt, sind insbesondere die konkreten Gen-Umwelt-Interaktionen im Nachhinein als oft sehr zweifelhaft zu betrachten und können häufig kaum repliziert werden. Gleichzeitig ist es eine der am besten gesicherten Erkenntnisse aus der Zwillingsforschung, dass so gut wie alle menschlichen Eigenschaften, psychologische wie physiologische, in ihrer Variation teilweise auf erblich bedingte Unterschiede zurückführbar sind (Polderman et al., 2015). Für die Diskrepanz zwischen dieser Tatsache und den geringen Erfolgen bei der Identifizierung konkreter genetischer Varianzquellen wurde der Begriff *missing heritability* geprägt (Maher, 2008; Young, 2019).

Analog zu dieser Entwicklung stellt sich die Lage in Bezug auf Umwelteinflüsse dar, insbesondere nichtgeteilte Umwelteinflüsse. Von wenigen Ausnahmen abgesehen, wie etwa bestimmten religiösen oder politischen Einstellungen (siehe Koenig et al., 2005; Hatemi et al., 2014), ist diese sogenannte „E“-Komponente in aller Regel deutlich einflussreicher als die mit „C“ bezeichnete Komponente der familiär geteilten Umweltvarianz (Polderman et al., 2015). Je nach konkreter statistischer Modellierung ist dies teilweise darauf zurückzuführen, dass in dieser Varianzkomponente auch sämtliche Fehlervarianz aufgeht, d.h. etwa die Unreliabilität von Fragebögen (Rijsdijk & Sham, 2002; van den Berg et al., 2007). Dennoch mündete die frühe Forschung zur nichtgeteilten Forschung bald in die Erkenntnis, dass sich parallel zur *missing heritability* auch ein *missing environment* konstatieren lässt (Røysamb & Tambs, 2016): konkrete Umwelteinflüsse zu identifizieren, die Teile der gefundenen nichtgeteilten Umweltvarianz erklären können, ist oft nicht erfolgreich. Bereits 2001 nannten Robert Plomin und Kollegen dies einen „düsteren Ausblick“ für die Verhaltensgenetik („gloomy prospect“; Plomin et al., 2001, S. 231). Ähnlich wie die Forschung zu genetischen Varianzquellen hat auch die Forschung zur nichtgeteilten Umwelt seitdem deutliche Fortschritte gemacht. Dennoch sind klare inhaltliche Durchbrüche, die große Teile des *missing environments* erklären würde, bis heute ausgeblieben – dies resümiert Plomin (2024) über zwei Jahrzehnte nach seinem ursprünglichen Essay. Auch andere Autoren konstatieren, dass nichtgeteilte Einflüsse so idiosynkratisch (Smith, 2011) und möglicherweise sogar auf der molekularen Ebene so zufällig (Tikhodeyev & Shcherbakova, 2019) auftreten könnten, dass eine Aufklärung inhaltlicher „E“-Variablen nur in kleinteiligen Schritten realistisch erscheint.

Diese Ausgangslage kann erklären, dass auch im Kontext immer komplexer werdender Ansätze zum Beispiel in der Molekulargenetik weiterhin klarer Bedarf an Forschung besteht, die „lediglich“ konkrete Umwelteinflüsse erfolgreich mit konkreten Outcomes verknüpft.

Das gilt insbesondere, als der eingangs erwähnte Status von eineiigen Zwillingspaaren als

„natürliches Experiment“ helfen kann, Konfundierungen zu vermeiden, welche mit anderen Methoden gewonnene Schlussfolgerungen möglicherweise verzerren. Neben den Modellen zur Varianzzerlegung von Eigenschaften ist ein weiteres wichtiges Standbein der Zwillingsforschung deshalb der direkte Blick auf die eineiigen Zwillinge als Paare von kontrafaktischen Duos. Jegliche Unterschiede innerhalb dieser Paare müssen bei gemeinsam aufwachsenden eineiigen Zwillingen auf nichtgeteilte Umwelteinflüsse zurückgehen (abgesehen von möglichen geringfügigen genetischen Unterschieden selbst bei MZ-Paaren; siehe Gringras & Chen, 2001). Dadurch sind sämtliche genetisch bedingte oder geteilt umweltbedingte Konfundierungen, die Vergleiche zwischen nicht genetisch identischen Personen verzerren können, kontrolliert (Carlin et al., 2005). Da ein solcher Standard in der Praxis nie durch explizites Miterfassen und Herausrechnen gleich wie vieler Kontrollvariablen erreichbar wäre, gilt die Analyse von Zwillingsunterschieden als einer der besten Ansätze, sich in einem vergleichsweise simplen Design kausalen Schlussfolgerungen anzunähern (Vitaro et al., 2009; McAdams et al., 2021). Tatsächliche Kausalität ist auch in diesem Paradigma allenfalls unter sehr spezifischen Voraussetzungen nachweisbar (Sjölander et al., 2012), zumal nichtgeteilt umweltbedingte Konfundierungen weiterhin nicht kontrolliert sind. Dennoch kann mit Vergleichen innerhalb von eineiigen Zwillingspaaren ein oft deutlich robusterer Nachweis für Zusammenhänge geleistet werden, die andernfalls zahlreichen Alternativerklärungen unterliegen würden (Johnson et al., 2009). Auf diese Weise konnten diese sogenannten *Twin-Difference-Modelle* beispielsweise helfen, den heute selbstverständlichen Forschungsstand zu etablieren, dass Rauchen gesundheitsschädlich ist (McGue et al., 2010).

Vergleiche innerhalb von Zwillingspaaren stellen somit oft eine effektive Art und Weise dar, korrelative Zusammenhänge belastbarer zu prüfen. Dabei finden sich nicht selten tatsächlich Nulleffekte für vermeintlich gut gesicherte oder sogar trivial wirkende Zusammenhänge, etwa zwischen negativen Lebensereignissen und schlechterem Schlaf (Barclay et al., 2012). Studie II nutzte diese Methode, um einen bis dahin nicht mithilfe von Zwillingsdaten überprüften vermuteten Zusammenhang zu testen, nämlich zwischen dem Besuch von Sportvereinen und höherer Lebenszufriedenheit.

Subjektive Lebenszufriedenheit

Subjektive Lebenszufriedenheit wird konzeptuell meist als Bestandteil des breiter gefassten Konstruktes *subjektives Wohlbefinden* eingeordnet, auch wenn über dessen genaue Definition kein vollständiger Konsens besteht (siehe Charlemagne-Badal et al., 2015). Neben der globalen Lebenszufriedenheit, bei der nach einer allgemeinen Einschätzung der eigenen Lebenssituation

gefragt wird, hat sich eine separate Betrachtung domänenspezifischer Zufriedenheit, etwa mit der eigenen beruflichen oder familiären Situation, als theoretisch wie empirisch sinnvoll erwiesen (Lent et al., 2005; Long & Huebner, 2014).

Die subjektive Lebenszufriedenheit enthält eine recht stabile, in der Persönlichkeit verankerte Komponente (Schimmack et al., 2002). Auch wenn die Schätzungen der Varianzanteile je nach Studie unterschiedlich ausfallen, können erblich bedingte Unterschiede insgesamt nur einen moderaten Anteil – im Durchschnitt 32% - in Unterschieden in der Lebenszufriedenheit aufklären (Bartels, 2015). Hahn und Kollegen (2013) stellten für diese erblich bedingten Varianzanteile eine nahezu vollständige Überlappung (genetische Korrelation) mit den genetisch bedingten Varianzanteilen in den Big Five fest, insbesondere Neurotizismus und Extraversion.

Diese vergleichsweise geringe Erblichkeit lässt viel Raum für umweltbedingte Varianzanteile und somit auch für die Suche nach konkreten Umwelteinflüssen, welche eine Rolle dabei spielen können, dass selbst genetisch identische Zwillinge mit ihrem jeweiligen Leben unterschiedlich zufrieden sind. Die umweltbedingten Varianzanteile an der Lebenszufriedenheit scheinen zudem überwiegend auf nichtgeteilte Faktoren zurückzugehen; nur in wenigen Studien (etwa Nes et al., 2008) findet sich Evidenz für geteilt umweltbedingte Komponenten. Damit ist die subjektive Lebenszufriedenheit ein Konstrukt, bei dem die Frage nach dem in Abschnitt 3 angesprochenen *missing environment* relevant ist.

In Studien mit dem Twin-Difference-Design ist nur teilweise festzustellen, dass mögliche Korrelate einer höheren Lebenszufriedenheit auch in diesem rigoroseren Ansatz bestehen bleiben. Mann und Kollegen (2019) konnten solche Zusammenhänge nur für wenige der von ihnen überprüften zahlreichen Konstrukte nachweisen, darunter insbesondere für soziale Unterstützung, aber nicht beispielsweise für den Erziehungsstil. Auch für Konstrukte wie Arbeitslosigkeit (Schnittker, 2008) oder die Big Five (Misheva, 2016) zeigen sich innerhalb eineiiger Zwillingspaare keine signifikanten Korrelationen mit der Lebenszufriedenheit mehr. Diener et al. (2018) weisen in ihrer Übersichtsarbeit darauf hin, dass zur Frage der Kausalität der gängigen Prädiktoren von Lebenszufriedenheit noch Forschungsbedarf bestehe. Die gemischten Ergebnisse in Twin-Difference-Studien unterstreichen dies.

Sportliche Aktivität und Lebenszufriedenheit

Studie II setzte bei einem weiteren Nulleffekt im Kontext von Twin-Difference-Studien zur Lebenszufriedenheit an: Stubbe und Kollegen (2007) betrachteten eineiige Zwillingspaare, bei denen jeweils ein Zwilling sportlich aktiv war und der andere nicht. Ein Unterschied in der

Lebenszufriedenheit der beiden Zwillinge war im Durchschnitt nicht feststellbar. Eine Replikation dieses Nullergebnisses über diese Einzelstudie hinaus wäre bemerkenswert, da Sport zu treiben allgemein als protektiver Faktor für das mentale Wohlbefinden gesehen wird (Buecker et al., 2021). Neben Studien, die körperliche Aktivität im Allgemeinen als Schutzfaktor für die mentale Gesundheit identifizieren (etwa Varca et al., 1984; Schmiedeberg & Schröder, 2017; Pietsch et al., 2022), zeigt weitere Forschung solche Effekte spezifisch für die Teilnahme an Sportvereinen im Gegensatz zu körperlicher Aktivität im Allgemeinen (Vilhjalmsson & Thorlindsson, 1992), insbesondere auch in Deutschland (Spengler & Woll, 2013; Mutz et al., 2021). Dies lässt sich möglicherweise auch auf den gemeinschaftlichen Aspekt solcher Vereine zurückführen, da diese die soziale Verbundenheit der Teilnehmenden stärken (Eime et al., 2010), welche einen der wichtigsten Prädiktoren höherer Lebenszufriedenheit darstellt (Becchetti et al., 2012). Stärkere körperliche Aktivität im Alltag ohne den Vereinsaspekt ist allgemein zwar offenbar auch mit höherem Wohlbefinden assoziiert (für einen Überblick siehe Maher & Conroy, 2017), hier ist die Befundlage jedoch weniger eindeutig (siehe etwa Pedišić et al., 2015; Wicker & Frick, 2015).

In diese insgesamt breite Evidenz für positive Assoziationen zwischen sportlicher Aktivität, insbesondere in einem Vereinsumfeld, und dem psychischen Wohlbefinden sowie spezifisch der Lebenszufriedenheit sind bisher kaum genetisch informierte Studien eingegangen. Dem Ergebnis von Stubbe und Kollegen (2007) steht eine Studie von Waller und Kollegen (2010) gegenüber, welche durchaus im Durchschnitt höhere Lebenszufriedenheits-Werte beim sportlich aktiveren Zwilling in diskordanten eineiigen Paaren fanden. Weitere Studien untersuchen ebenfalls eineiige Zwillingspaare, die im Hinblick auf körperliche Aktivität diskordant sind, fokussieren sich dabei aber auf andere Aspekte der Lebensqualität (Kujala et al., 1998; Leskinen et al., 2009; Korhonen et al., 2009). Spezifisch für die Teilnahme an Sportvereinen im Gegensatz zu allgemeiner sportlicher Aktivität sind keine genetisch informierte Studien bekannt.

Studie II wurde deshalb mit dem Ziel durchgeführt, in einer genetisch informierten Stichprobe zu überprüfen, ob der mit anderen Methoden wiederholt gefundene positive Zusammenhang zwischen dem Besuch von Sportvereinen und subjektiver Lebenszufriedenheit über genetische und geteilt umweltbedingte Konfundierungen hinaus abgesichert werden kann.

Ergebnisse und Schlussfolgerungen

Für Studie II wurden die Alterskohorten 3 und 4 des TwinLife-Panels herangezogen. Entsprechend der hohen gesellschaftlichen Bedeutung von Sportvereinen in Deutschland (Breuer et al., 2015) besuchten in Kohorte 3 über die Hälfte und in Kohorte 4 noch rund 38% der Zwillinge regelmäßig

einen Sportverein. Als „regelmäßiger Besuch“ wurde dabei ein mindestens wöchentlicher Besuch aufgefasst.⁴ Auch in Kohorte 2 liegen (hier eltern- statt selbstberichtete) Angaben über den Besuch von Sportvereinen vor, jedoch sind diskordante Paare hier insbesondere bei den eineiigen Zwillingen mit einer Häufigkeit von nur 4% zu selten, als dass ein Einbezug dieser deutlich jüngeren Kinder in die Gesamtstichprobe für eine Twin-Difference-Analyse als sinnvoll erachtet wurde. In den älteren Kohorten steigen die Diskordanzraten an, auch wenn diskordante Paare gerade bei den eineiigen Zwillingen auch hier in der Minderheit bleiben: in Kohorte 3 sind 14.4% der eineiigen und 28.2% der zweieiigen Zwillingspaare diskordant, in Kohorte 4 steigen diese Werte auf 19.8% beziehungsweise 33.2%. Die innerhalb jeder Alterskohorte höhere Diskordanz innerhalb zweieiiger als innerhalb eineiiger Zwillingspaare deutet darauf hin, dass auch erblich bedingte Faktoren Varianz im Besuch von Sportvereinen aufklären. Dies entspricht früheren Befunden, nach denen die Varianz in sportlichen Aktivitäten in der Freizeit eine substantielle erbliche Komponente aufweist (Carlsson et al., 2006).

Globale Lebenszufriedenheit wurde in TwinLife mit der fünf Items umfassenden *Satisfaction with Life Scale* abgefragt (SWLS; Diener et al., 1985). Zudem wurden auf 11-stufigen Einzelitems die domänenspezifische Zufriedenheit mit verschiedenen Lebensbereichen erfragt. Für Studie II wurden die Fragen nach der Zufriedenheit mit der eigenen Freizeitgestaltung, der eigenen Gesundheit und dem eigenen Freundeskreis berücksichtigt.

Tabelle 3 gibt die Parameter aus univariaten Zwillingenmodellen für diese Lebenszufriedenheits-Maße wieder (modelliert mit dem R-Package *umx*; Bates et al., 2019). Dabei zeigte in einigen Fällen ein ADE-Modell, also eines mit Berücksichtigung dominanter erblicher Effekte (in Ergänzung zu rein additiven) auf Kosten einer nicht modellierten geteilten Umweltkomponente, einen besseren Fit zu den Daten. Insgesamt sind die Modelle jedoch vor allem von deutlichen nichtgeteilten Umweltkomponenten gekennzeichnet. Dies ist mutmaßlich teilweise auf die verwendeten sehr kurzen und damit unreliablen Maße zurückzuführen, die eine hohe Fehlervarianz mit sich bringen. Die genetischen Varianzanteile liegen mit zwischen 16% und 31% im unteren Bereich der Spanne bisheriger Untersuchungen (siehe Bartels, 2015). Globale Lebenszufriedenheit zu beiden Zeitpunkten sowie Zufriedenheit mit der eigenen Freizeitgestaltung zu Zeitpunkt 2 beinhalteten substantielle geteilte Umweltanteile.

4 Eine überwiegende Mehrheit der Teilnehmer (92.5%) gaben entweder einen solchen wöchentlichen Besuch oder aber keine Aktivität in Sportvereinen an; die übrigen Fälle, die einen selteneren Besuch angaben, wurden der nicht in Sportvereinen aktiven Gruppe zugeordnet. Ein *sensitivity check* zeigt, dass die Ergebnisse der weiteren Analysen sich nicht bedeutsam ändern, wenn diese Kategorien nicht zusammengefasst werden.

Tabelle 3. Geschätzte Varianzanteile der Lebenszufriedenheits-Maße in univariaten Zwillingsmodellen

	A	C	E	D
Globale Lebenszufriedenheit (Zeitpunkt 1)	.21 [.08, .36]	.20 [.08, .32]	.59 [.54, .65]	-
Domänenspezifische Zufriedenheit: Gesundheitszustand (Zeitpunkt 1)	-	-	.72 [.66, .79]	.28 [.21, .34]
Domänenspezifische Zufriedenheit: Freizeitgestaltung (Zeitpunkt 1)	.22 [.16, .27]	-	.78 [.73, .84]	-
Domänenspezifische Zufriedenheit: Freundschaften (Zeitpunkt 1)	-	-	.69 [.63, .75]	.31 [.25, .37]
Globale Lebenszufriedenheit (Zeitpunkt 2)	-	.35 [.29, .41]	.65 [.59, .71]	-
Domänenspezifische Zufriedenheit: Gesundheitszustand (Zeitpunkt 2)	-	-	.84 [.77, .90]	.16 [.14, .23]
Domänenspezifische Zufriedenheit: Freizeitgestaltung (Zeitpunkt 2)	-	.16 [.14, .23]	.84 [.77, .90]	-
Domänenspezifische Zufriedenheit: Freundschaften (Zeitpunkt 2)	-	-	.78 [.70, .87]	.22 [.13, .30]

Anmerkung. Dargestellt ist jeweils das Modell mit der besten statistischen Anpassungsgüte. 95%-Konfidenzintervalle in Klammern.

Zur Modellierung der Twin-Difference-Analysen kann folgende Regressionsgleichung genutzt werden (siehe Carlin et al., 2005):

$$E(Y_{ij}) = \beta_0 + \beta_w * (X_{ij} - \bar{X}_i) + \beta_B * \bar{X}_i$$

Die sogenannte *between*-Komponente (β_B) stellt dabei die Abweichung des Durchschnittswerts eines Zwillingspaars im Kriterium vom Durchschnitt der gesamten Stichprobe dar, die *within*-Komponente (β_w) die Abweichung eines Zwillings vom Durchschnittswert des jeweiligen Paares (d.h. die Hälfte der Differenz zwischen der betreffenden Person und ihrem Zwilling). Im Falle eines dichotomen Outcomes, wie in Studie II gegeben, kontrastiert β_B dabei konkordante betroffene Paare (d.h., beide Zwillinge besuchen einen Sportverein) mit konkordanten nicht betroffenen Paaren (keiner der Zwillinge besucht einen Sportverein), und der Parameter β_w kontrastiert betroffene und nicht betroffene Zwillinge innerhalb der diskordanten Paare (Arsenault et al., 2008). Das Modell geht somit in ein Co-Twin-Design über (siehe McAdams et al., 2021; Vitaro et al., 2009).

Für die Identifikation der interessierenden Effekte über genetische und geteilt umweltbedingte Konfundierungen hinaus ist der Vergleich innerhalb der eineiigen Zwillingspaare im Prinzip ausreichend. Die gleichzeitige Modellierung von within- und between-Komponenten als sogenannte *mixed strategy* hat sich jedoch mit der Zeit als häufigerer Ansatz etabliert, da der Vergleich der beiden Komponenten eine genauere Interpretation ermöglicht, in welchem Ausmaß die phänotypischen Zusammenhänge konfundiert sind (Carlin et al., 2005; Vitaro et al., 2009). Wenn keinerlei solche Konfundierung existiert, liegt der Erwartungswert für das

Regressionsgewicht der between-Komponente bei 0, da alle Zusammenhänge in diesem Fall in der within-Komponente abgebildet werden. Ein Scheinzusammenhang, der ausschließlich die Folge genetischer oder geteilt umweltbedingter Konfundierung ist, sollte sich umgekehrt darin zeigen, dass nur die between-Komponente signifikant von 0 abweicht. Signifikante Beiträge beider Komponenten sind entsprechend als gemischte Evidenz dafür zu interpretieren, dass die Ergebnisse teilweise, aber nicht vollständig konfundiert sind (Carlin et al., 2005).

Aufgrund der nicht gegebenen Unabhängigkeit der Daten innerhalb von Zwillingspaaren wurde die Zugehörigkeit zu einem bestimmten Zwillingspaar als random effect in das Regressionsmodell aufgenommen (siehe Carlin et al., 2005) sowie für die innerhalb der (in TwinLife immer gleichgeschlechtlichen) Zwillingspaare notwendigerweise identischen Faktoren Alter und Geschlecht kontrolliert (McGue & Bouchard, 1984). Mit Ausnahme der dichotomen Variable zum Besuch von Sportvereinen gingen alle Variablen in z-standardisierter Form in die Analyse ein. Die Angaben zur Lebenszufriedenheit wiesen eine tendenziell rechtssteile Verteilung auf, wie dies bei Fragen nach der Lebenszufriedenheit häufig der Fall ist (Pavot & Diener, 2009). Empfehlungen in der statistischen Literatur entsprechend (siehe Schielzeth et al., 2020; Schmidt & Finan, 2018) wurde von einer Transformation dieser Verteilung abgesehen.

Tabelle 4. Regressionsgewichte aus linearen gemischten Modellen zur Vorhersage der globalen und domänenspezifischen Lebenszufriedenheit aus Unterschieden in der Teilnahme an Sportvereinen innerhalb und zwischen Zwillingspaaren sowie Alter und Geschlecht

Kriterium	Prädiktor	Gesamte Stichprobe (n = 877 MZ-Paare, 869 DZ-Paare)	Nur MZ-Paare (n = 877 Paare)	Nur MZ-Paare, Längsschnitt (Kriterium zu Zeitpunkt 2 gemessen; n >= 899 Personen, siehe Anmerkungen)
GLS	age	0.03 [-0.04, -0.01]**	-0.01 [-0.02, 0.01]	-0.01 [-0.01, 0.02]
	sex	-0.09 [-0.19, 0.02]	0.07 [-0.03, 0.18]	-0.03 [-0.13, 0.08]
	β_B	0.32 [0.23, 0.40]***	0.27 [0.15, 0.38]***	0.07 [-0.05, 0.19]
	β_W	0.15 [0.04, 0.25]**	0.01 [-0.16, 0.18]	0.05 [-0.18, 0.27]
	β_{Bsat}			0.75 [0.69, 0.82]***
	β_{Wsat}			0.47 [0.38, 0.56]***
DLS: Gesundheitszustand	age	-0.02 [-0.03, -0.01]***	-0.03 [-0.04, -0.01]**	-0.00 [-0.02, 0.02]
	sex	-0.06 [-0.13, 0.01]	-0.09 [-0.19, 0.02]	-0.03 [-0.15, 0.10]
	β_B	0.12 [0.04, 0.20]**	0.11 [0.00, 0.22]	0.03 [-0.11, 0.16]
	β_W	0.26 [0.14, 0.38]***	0.31 [0.13, 0.49]***	0.09 [-0.16, 0.35]
	β_{Bsat}			0.55 [0.47, 0.63]***

	β_{Wsat}			0.34 [0.25, 0.43]***
DLS:	age	-0.02 [-0.03, 0.01]**	-0.02 [-0.17, 0.03]*	-0.00 [-0.02, 0.02]
Freizeitgestaltung	sex	-0.05 [-0.12, 0.02]	-0.07 [-0.17, 0.03]	-0.05 [-0.18, 0.08]
	β_B	0.20 [0.12, 0.29]***	0.23 [0.11, 0.34]***	0.14 [-0.01, 0.28]
	β_W	0.19 [0.07, 0.31]***	0.21 [0.02, 0.40]*	-0.08 [-0.35, 0.19]
	β_{Bsat}			0.41 [0.33, 0.50]***
	β_{Wsat}			0.29 [0.19, 0.38]***
DLS:	age	-0.02 [-0.03, 0.00]**	-0.01 [-0.03, 0.00]	-0.00 [-0.03, 0.02]
Freundschaften	sex	0.01 [-0.06, 0.07]	-0.00 [-0.11, 0.10]	-0.02 [-0.15, 0.12]
	β_B	0.12 [0.04, 0.20]**	0.10 [-0.01, 0.21]	0.10 [-0.05, 0.25]
	β_W	0.19 [0.08, 0.31]**	0.17 [-0.01, 0.34]	0.05 [-0.22, 0.31]
	β_{Bsat}			0.46 [0.36, 0.55]***
	β_{Wsat}			0.35 [0.24, 0.45]***

Anmerkungen: β_B = Teilnahme eines Zwillingspaars an Sportvereinen im Vergleich zum stichprobenweiten Mittelwert der Teilnahme (between-pair effect); β_W = Teilnahme eines Zwillings an Sportvereinen im Vergleich zum jeweils anderen Zwilling des Paares (within-pair effect). Analog dazu in längsschnittlichen Analysen: β_{Bsat} = between-pair effect für globale bzw. domänenspezifische Lebenszufriedenheit zu Zeitpunkt 1; β_{Wsat} = within-pair effect für globale bzw. domänenspezifische Lebenszufriedenheit zu Zeitpunkt 1. GLS = globale Lebenszufriedenheit; DLS = domänenspezifische Lebenszufriedenheit; DZ-Paare = zweieiige Zwillingspaare; MZ-Paare = eineiige Zwillingspaare. Intercepts aus Gründen der Übersichtlichkeit nicht berichtet. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$. Stichprobengrößen der längsschnittlichen Analysen (aufgrund verschiedentlich auftretender fehlender Werte): $n = 899$ Personen für GLS, 945 für DLS - Gesundheitszustand, 932 für DLS - Freizeitgestaltung, 927 für DLS – Freundschaften.

Die Ergebnisse in Tabelle 4 zeigen innerhalb der eineiigen Zwillingspaare ein gemischtes Bild. Bei der globalen Lebenszufriedenheit ist die within-Komponente nicht signifikant von Null verschieden. Der korrelative Zusammenhang zwischen dem Besuch von Sportvereinen und der globalen Lebenszufriedenheit könnte somit aufgrund einer genetischen oder geteilt umweltbedingten Konfundierung beider Variablen zustandekommen. Dies bestätigt die früheren Ergebnisse von Stubbe und Kollegen (2007), in denen nur die globale Lebenszufriedenheit betrachtet wurde. Da die within-Komponente in der vollen Stichprobe, also inklusive der zweieiigen Zwillingspaare, dagegen signifikant positiv ist, ist eine genetische Konfundierung die plausiblere Erklärung als eine geteilt umweltbedingte Konfundierung (siehe Carlin et al., 2005). Für zwei der drei domänenspezifischen Items, nämlich jene zur Zufriedenheit mit der eigenen Gesundheit und mit der eigenen Freizeitgestaltung, bleibt die within-Komponente dagegen auch

innerhalb der eineiigen Zwillingspaare signifikant und sinkt numerisch im Vergleich zur vollen Stichprobe auch nicht ab. Diese Zusammenhänge sind somit in den aktuellen Daten nicht durch genetische (oder geteilte umweltbedingte) Konfundierung erklärbar, sondern deuten auf eine zumindest zum Teil nichtgeteilt umweltbedingte Vermittlung hin. Dies gilt insbesondere für die Zufriedenheit mit der eigenen Gesundheit, bei der die between-Komponente nicht signifikant von 0 verschieden ist (wenn auch angesichts des Konfidenzintervalls sehr knapp).

Bei der Frage nach dem eigenen Freundeskreis erreicht innerhalb der eineiigen Zwillingspaare keine der beiden Komponenten statistische Signifikanz, wobei zu beachten ist, dass die Regressionsgewichte deskriptiv ähnlich hoch ausfallen wie in der gesamten Stichprobe, jedoch aufgrund der ungefähr halbierten Stichprobe (Betrachtung nur noch der eineiigen Zwillingspaare) mit höheren Unsicherheitsbereichen behaftet sind.

Die signifikanten within-Komponenten bei eineiigen Zwillingspaaren für zwei Aspekte der domänenspezifischen Lebenszufriedenheit spiegeln sich darin wieder, dass innerhalb dieser Paare derjenige Zwilling, der regelmäßig einen Sportverein besucht, im Durchschnitt eine höhere Zufriedenheit mit der eigenen Gesundheit und Freizeitgestaltung berichtet. Deskriptiv liegen die Werte auf den 11-stufigen Einzelitems, mit denen die domänenspezifische Lebenszufriedenheit abgefragt wurde, für die Zufriedenheit mit der eigenen Gesundheit bei 8.30 (SD = 1.53) bei den sporttreibenden eineiigen Zwillingen und bei 7.58 (SD = 2.39) bei ihren diskordanten Zwillingen. Für die Zufriedenheit mit der eigenen Freizeitgestaltung betragen die Mittelwerte 7.85 (SD = 1.92) beziehungsweise 7.36 (SD = 2.32). Diese Unterschiede sind in t-Tests nach Welch (siehe Declare et al., 2017) mit $t(255.55) = 3.09$, $p = 0.002$ sowie mit $t(292.40) = 1.93$, $p = 0.4949$ signifikant. Es handelt sich jeweils um kleine Effekte (Cohens $d = 0.36$ beziehungsweise 0.23). Bei der Zufriedenheit mit dem eigenen Freundeskreis und bei der globalen Lebenszufriedenheit weisen die Zwillinge in diskordanten eineiigen Paaren, die einen Sportverein besuchen, dagegen keine signifikant höheren Zufriedenheitswerte auf als die nicht aktiven Zwillinge in diesen Paaren. Hier sind solche deskriptiven Unterschiede nur innerhalb der zweieiigen Zwillingspaare signifikant ($t(524.78) = 2.61$, $p = 0.009$ für die globale Lebenszufriedenheit, $t(482.27) = 2.59$, $p = 0.010$ für die Zufriedenheit mit dem Freundeskreis). Dies bestätigt die Schlussfolgerung aus Tabelle 4, dass diese Unterschiede nicht über eine Kontrolle genetischer und geteilt umweltbedingter Konfundierung hinaus bestehen bleiben.

Keine der genannten Zusammenhänge bleiben bestehen, wenn die Lebenszufriedenheit zum Zeitpunkt der zweiten TwinLife-Befragungswelle als Kriterium herangezogen wird und für die Lebenszufriedenheit zu Zeitpunkt 1 kontrolliert wird (siehe Tabelle 4). Auch wenn die deutlich verringerte Teilstichprobe eine gesicherte Einordnung dieser Ergebnisse erschwert, deutet dieses

Muster darauf hin, dass die querschnittlichen Ergebnisse durch das Ausgangslevel des Kriteriums konfundiert sein könnten. Mit anderen Worten könnten Unterschiede in der früheren Lebenszufriedenheit erklären, warum ein Zwilling später einen Sportverein besucht und der andere nicht, und damit einen Scheinzusammenhang zwischen dem Vereinsbesuch und der aktuellen Lebenszufriedenheit herstellen. Um diese mögliche Alternativerklärung abschließend zu überprüfen, wäre eine größere Stichprobe mit höherer längsschnittlicher Stabilität erforderlich. Dies unterstreicht, dass die gefundenen querschnittlichen Unterschiede zwischen Zwillingspaaren nur eingeschränkt als Hinweis auf tatsächliche Kausalzusammenhänge zwischen dem Besuch von Sportvereinen und einer höheren domänenspezifischen Lebenszufriedenheit mit Gesundheit und Freizeitgestaltung aufgefasst werden können. Nur genetisch und geteilt umweltbedingt erklärbare Varianzanteile sind als konfundierende Variablen kontrolliert, nicht aber nichtgeteilt umweltbedingte. Dazu kommen verschiedene Limitationen der Stichprobe und Variablen, die sich möglicherweise jedoch auch in Richtung einer möglichen Unterschätzung der Effekte ausgewirkt haben könnten. Allen voran ist von den Zwillingen in TwinLife nur bekannt, *ob* sie einen Sportverein regelmäßig besuchen, aber keine weiteren Begleitumstände, die eine entscheidende Rolle für mögliche positive Effekte spielen könnten: die Beweggründe für das Besuchen des Vereins, die konkrete Sportart, das soziale Umfeld im Verein sind nur einige mögliche Komponenten. Dazu kommt die ebenfalls sehr grobkörnige Messung der Lebenszufriedenheit, im Falle der domänenspezifischen Zufriedenheit mit Einzelitems. Auch die sehr ausgeprägten nichtgeteilt umweltbedingten Varianzkomponenten in den univariaten genetischen Modellen sprechen für ein hohes Ausmaß an Messungenauigkeit.

Zu den Ergebnissen in Tabelle 4 lässt sich des Weiteren feststellen, dass in der längsschnittlichen Betrachtung die within-Komponenten der Lebenszufriedenheit zu Zeitpunkt 1 weiterhin bedeutende Anteile an der Vorhersage der Lebenszufriedenheit zu Zeitpunkt 2 zeigen. Auch wenn diese Koeffizienten geringer ausfallen als die between-Komponenten, deutet dies dennoch darauf hin, dass ein substantieller Anteil in den *Unterschieden* in der Lebenszufriedenheit innerhalb der eineiigen Zwillingspaare zeitlich stabil ist und somit systematische Unterschiede zwischen den Zwillingen in eineiigen Paaren widerspiegelt, welche auf nichtgeteilte Umwelteinflüsse zurückzuführen sein müssen. Die Suche nach dem *missing environment* ist hier also weiterhin sinnvoll, auch wenn eine solche Stabilität nicht ausschließen kann, dass es sich um eine Anhäufung einzelner sehr kleiner, schwierig zu identifizierender Effekte handeln könnte.

4 Labeling-Effekte im Kontext psychischer Erkrankungen

Den zweiten Teil der zur vorliegenden Dissertation gehörenden Forschungsvorhaben stellten Erhebungen zur Stigmatisierung psychischer Erkrankungen dar. Menschen mit psychischen Erkrankungen befinden sich häufig in einer Situation, die mehrere der eingangs (siehe Abschnitt 1) erwähnten Aspekte subjektiver Lebensqualität beeinträchtigt (Corrigan, 2005). In Form einer doppelten Belastung wird neben dem unmittelbaren Leidensdruck durch die Erkrankung häufig auch der Bereich der sozialen Integration in Mitleidenschaft gezogen, da die Betroffenen Stigmatisierung und Ausgrenzung erleben (Pescosolido et al., 2013; Schomerus et al., 2023). Dies kann sich konkret etwa in geringeren Chancen auf dem Wohnungsmarkt zeigen (Flage & LeGallo, 2025) und so die materielle Unabhängigkeit als einen der von der Weltgesundheitsorganisation (WHOQoL SRPB Group, 2006) definierten Aspekte der Lebensqualität beeinträchtigen. Durch ein Internalisieren der erfahrenen Stigmatisierung kann sich in Form von Selbst-Stigma ein Teufelskreis entwickeln, da die Betroffenen beispielsweise zögern, sich Hilfe zu suchen (Schnyder et al., 2017).

Zu diesem Thema wurde neben verschiedenen studentischen Projekten und Abschlussarbeiten auch Studie III (Dings & Kubillus, 2025) durchgeführt und publiziert, die im Folgenden vorgestellt werden soll.

Diagnostische Label als mögliche Treiber von Stigmatisierung

Auch wenn bisher kein einzelnes theoretisches Modell zur Erklärung der Stigmatisierung psychischer Erkrankungen als etablierter Konsens gelten kann, wird diese häufig auf Basis allgemeiner sozialpsychologischen Theorien zu Integration und Ausgrenzung als sozialer Etikettierungsprozess beschreiben (siehe Fox et al., 2018; Rüsch et al., 2020). Vorurteile gegenüber bestimmten Personen führen demnach dazu, dass diese als Teil einer mit bestimmten negativen Etiketten belegten Outgroup wahrgenommen werden. Konkret gelten Betroffene psychischer Erkrankungen beispielsweise als weniger kompetent oder, je nach Diagnose, als gefährlich (Rüsch et al., 2005; Jorm et al., 2012). Diese Etikettierung führt zu einer größeren sozialen Distanz (siehe Bogardus, 1926; Jorm & Oh, 2009) und passiver wie aktiver Ausgrenzung und Schädigung der Betroffenen (Görzig & Ryan, 2022). Letzterer Prozess wird über Emotionen wie Furcht, Wut oder Herablassung gegenüber den Betroffenen vermittelt (Sadler et al., 2015).

Zu den vielen offenen Forschungsfragen im Kontext der Stigmatisierung psychischer Erkrankungen zählt diejenige, welche Rolle das sprachliche Framing psychischer Erkrankungen und der

betroffenen Menschen in der Entstehung stigmatisierender Einstellungen einnimmt. In dem beschriebenen prototypischen Prozess hin von Vorurteilen zu tatsächlicher Ausgrenzung betrifft dies den Schritt der „Etikettierung“ der Betroffenen mit bestimmten Begriffen. Vergleichbar mit der Sapir-Whorf-Hypothese, welche Sprache nicht nur als neutrales Mittel der Kommunikation ansieht, sondern ihr umgekehrt auch aktiven Einfluss auf die Kognition zuschreibt (Sapir, 1929), wird in der Literatur häufig angenommen, dass der Sprachgebrauch im Zusammenhang mit psychischen Erkrankungen unmittelbar zur Entstehung von Stigma beitragen könnte (Rüsch et al., 2005; Volkow et al., 2021). Empirisch besteht jedoch weniger Klarheit, ob und auf welche Weise unterschiedlicher Sprachgebrauch tatsächlich die Einstellungen gegenüber Menschen mit psychischen Erkrankungen prägt. So sind etwa die Effekte sogenannter *person-first language*, welche Betroffene als Menschen „mit“ einer bestimmten Erkrankung bezeichnet („eine Person mit Schizophrenie“), statt sie mit der Erkrankung sprachlich gleichzusetzen (jemand „ist schizophren“), weiterhin umstritten (Granello & Gibbs, 2016; Masland & Null, 2022). Auch ob eine psychische Erkrankung als „Krankheit“ („disorder“) bezeichnet oder mit „milderen“ Begriffen benannt wird, zeigte in einer umfangreichen experimentellen Studie keine Effekte auf stigmatisierende Einstellungen (Fox et al., 2021).

Studie III befasste sich mit einer weiteren Form solcher vermuteten sprachlichen Framing-Effekte: der Benennung einer psychischen Erkrankung in Form eines diagnostischen Labels im Gegensatz zum Vermeiden einer solchen klinischen „Etikettierung“ einer Person.

Den historischen Ausgangspunkt der Forschung zu Labeling-Effekten bildet das Buch *Being mentally ill: A sociological theory* (Scheff, 1966), in dem der Soziologe Thomas Scheff diagnostische Label nicht nur als verstärkenden Faktor bei der Stigmatisierung psychischer Erkrankungen vermutete, sondern als eine zentrale Ursache dafür postulierte. Nach einer umfangreichen akademischen Debatte (siehe Gove, 1970; Horwitz, 1979) bildete sich eine weniger radikale *modified labeling theory* (Link et al., 1989) als Kompromiss heraus, nach welcher Label psychische Erkrankungen zwar nicht ursächlich ausmachen, aber durchaus zu deren stigmatisierender Abgrenzung vom „Normalen“ beitragen.

Obwohl diese Kernannahme der *modified labeling theory* so breite positive Rezeption gefunden hat, dass Imhoff (2016) ihr den Status einer „Binsenweisheit“ („cultural truism“; S. 461) zuschreibt, ist ihr empirischer Status weniger eindeutig. O'Connor und Kollegen (2021) kamen in einer Meta-Analyse zu dem Schluss, dass die Ergebnisse der existierenden experimentellen Studien, bei denen Fallvignetten mit oder ohne Nennung einer Diagnose präsentiert wurden, zu heterogen für kategorische Aussagen seien: „labeling effects cannot meaningfully be discussed in general 'catch-all' terms“ (S. 7).

Für Betroffene ist die Frage danach, ob die Nennung einer Diagnose Labeling-Effekte nach sich

zieht und welcher Art diese Effekte sind, oft von hoher Relevanz. Es liegen gemischte Erkenntnisse dazu vor, ob und wann ein Offenlegen einer psychischen Erkrankung empfehlenswert ist (Jones, 2011; Hastuti & Timmings, 2021; Pheister et al., 2020; Zamir et al., 2022). Vor diesem Hintergrund sind belastbare und insbesondere auch ökologisch valide Erkenntnisse zu Labeling-Effekten von hohem Wert für die Praxis.

Labeling-Effekte auf der Ebene einzelner Verhaltensweisen

Der Ausgangspunkt zur Durchführung von Studie III lag darin, dass es als unklar angesehen werden kann, ob die bisherigen Studien diesen Anspruch ökologischer Validität tatsächlich leisten können. Wie O'Connor und Kollegen (2021) sowie Imhoff (2016) ausführen, sind einige Studien zu dem Thema aufgrund grundsätzlicher Konfundierungen nur schwierig interpretierbar. So variieren zahlreiche Studien zum Thema die Nennung des Labels nicht experimentell, sondern definieren Labeling stattdessen dadurch, ob die Versuchspersonen von sich aus beispielsweise einer Fallgeschichte eine bestimmte Diagnose zuordnen oder nicht (für einen Überblick siehe Jorm & Oh, 2009). Solche „inferred labels“ (Manago & Mize, 2024) riskieren jedoch eine Konfundierung mit Merkmalen der beurteilenden Person, etwa ihrem Vorwissen über psychische Erkrankungen, und sind somit nicht kausal interpretierbar (Imhoff, 2016). Ähnliche Validitätsprobleme weisen Studien auf, bei denen eine als psychisch erkrankt etikettierte fiktive Person mit einer Person mit einem anderen Label – etwa einer körperlichen Erkrankung – oder auch mit einer neutralen Kontrollbedingung, also einer Person ohne Symptome psychischer Erkrankungen, verglichen wird. Eventuelle Unterschiede in einem solchen Vergleich können nicht auf die bloße An- oder Abwesenheit des Labels „psychische Erkrankung“ zurückgeführt werden (Imhoff, 2016). Darüber hinaus ist in der bisherigen Forschung festzustellen, dass auch fast alle Studien, welche diese Probleme nicht aufweisen, das gleiche Paradigma zur Erforschung von Labeling-Effekten nutzen, nämlich Fallvignetten. Dass mit der Beschränkung auf diesen Ansatz möglicherweise nicht vollständig erfasst wird, wie Labeling-Effekte in der Praxis auftreten, deutet eine Studie von Martinez und Kollegen (2011) an. Während ein globales Label „psychische Erkrankung“ hier negativere Bewertungen nach sich zog als das Vergleichslabel „körperliche Erkrankung“, zeigten sich umgekehrt *positive* Labeling-Effekte, wenn eine konkrete Situation geschildert wurde, in der eine Person ambivalentes, aber sozial unerwünschtes Verhalten zeigte. Eine durch Nennung der Diagnose klargestellte psychische Erkrankung sorgte hier für eine weniger verurteilende Interpretation des Verhaltens. Vor dem Hintergrund dieser Kontextabhängigkeit von Labeling-Effekten erscheint es angezeigt, diese nicht ausschließlich im Rahmen allgemein formulierter

Fallvignetten zu untersuchen. Es fehlen Erkenntnisse dazu, wie die Nennung einer Diagnose die Bewertung von isolierten Verhaltensweisen beeinflusst, bei denen keine umfassenderen Informationen über die betreffende Person gegeben sind.

Zudem konzentrieren sich die meisten Studien zu Labeling-Effekten auf eine globale Einschätzung der Person (O'Connor et al., 2021). Für Betroffene könnte es auch von Interesse sein, welche spezifischen Bewertungskategorien eventuellen Labeling-Effekten unterliegen, ob ein bestimmtes Verhalten unter Nennung eines diagnostischen Labels also beispielsweise als ungewöhnlicher eingeschätzt wird, ohne dass damit notwendigerweise eine Valenz einhergeht.

In Studie III sollte somit untersucht werden, ob und in welcher Hinsicht Verhaltensweisen, die tatsächlich diagnostisch relevante Symptome bestimmter psychischer Erkrankungen abbilden, in Anwesenheit versus Abwesenheit eines diagnostischen Etiketts unterschiedlich wahrgenommen und bewertet werden, auch wenn sie für sich stehend statt im Kontext einer Fallvignette präsentiert werden.

Methoden

Dazu wurden Items formuliert, die jeweils eine konkrete, möglichst umgrenzte Symptomatik abbildeten, beispielsweise für Schizophrenie: „Jemand hört innerliche Stimmen, die sie/ihn verurteilen oder Handlungen befehlen.“ Basis für die Konstruktion dieser Items war das DSM-5 in der zum Studienzeitpunkt vorliegenden Version (American Psychiatric Association, 2013). Dabei wurden im Vorfeld verschiedene Kriterien festgelegt, welche die Items nach Möglichkeit erfüllen sollten:

1. *Spezifität.* Die Symptome verschiedener psychischer Erkrankungen überlappen teils stark (Forbes et al., 2024). Soweit wie dennoch möglich, wurde angestrebt, Verhaltensweisen auszuwählen, die für die jeweilige Erkrankung spezifisch sind, damit die Präsentation eines Items mit Label keine inhaltliche Eingrenzung auf eine bestimmte Diagnose bewirkt, während gleichzeitig das Item ohne Label mehreren verschiedenen Diagnosen zugeordnet werden könnte.
2. *Alltäglichkeit.* Die Vermutung, dass die Anwesenheit eines Labels erst eine ansonsten nicht vorhandene negative, pathologisierende Wahrnehmung hervorruft, impliziert, dass das genannte Verhalten ohne ein Label alltäglich genug klingen muss, dass es als „normal“ wahrgenommen werden könnte. Tritt eine bestimmte, ungewöhnliche Verhaltensweise überhaupt nur im Rahmen einer bestimmten psychischen Erkrankung auf, wäre anzunehmen, dass bei entsprechendem Fachwissen dieses charakteristische Verhalten ein

Label auch ohne Nennung zumindest evozieren könnte. Es wurde somit versucht, die Items auf solche Symptome zu beschränken, deren Interpretationsspielraum auch eine Deutung als alltägliche, nicht pathologische Verhaltensweise zulässt.

3. *Unidirektionalität*. Nach Möglichkeit wurden Symptome vermieden, die in gegenteilige Richtungen auftreten können (etwa ein veränderter Schlafbedarf, der im Rahmen depressiver Erkrankungen sowohl erhöht als auch verringert sein kann).

Zusätzlich zu den Items, die tatsächliche Diagnosekriterien im DSM-5 widerspiegeln, wurden Verhaltensweisen formuliert, die tatsächlichen Symptomen ähneln oder populärwissenschaftlich häufig dafür gehalten werden (beispielsweise eine vermeintliche „gespaltene Persönlichkeit“ bei Schizophrenie), aber im DSM-5 nicht tatsächlich zur Beschreibung der jeweiligen Diagnose genutzt werden. Beispielsweise lautete ein solches „Pseudo-Item“ für Sozialphobie: „Jemand reagiert verunsichert auf negative Rückmeldung und Kritik.“

Ziel dieser Items war eine weitergehende Überprüfung der Annahme, dass ein Label (teilweise) zur Wahrnehmung einer Verhaltensweise als pathologisch oder andersartig beiträgt. Ein solcher ursächlicher Effekt wäre überzeugender nachgewiesen, wenn er unabhängig von der tatsächlichen Aussagekraft der Verhaltensweise aufträte. Im Rahmen von Studien mit Fallvignetten konnte bereits gezeigt werden, dass auch falsche Label („Schizophrenie“ bei Depressions-Symptomen und umgekehrt; Matsunaga & Kitamura, 2016) sowie im Laufe der Studie als Fehlinformation revidierte Label (Mickelberg et al., 2024) dennoch mit Labeling-Effekten einhergehen. Die Pseudo-Items dienten somit der Überprüfung, ob auch im hier gewählten Ansatz auf Itemebene inhaltlich nicht valide Label weiterhin Labeling-Effekte evozieren.

Die Items in Studie III nannten Symptome fünf psychischer Erkrankungen: „F32.1 Major Depression“ (6 Symptome / 2 Pseudo-Items), „F40.10 Soziale Angststörung“ (4 / 1), „F20.9 Schizophrenie“ (3 / 1), „F50.01 Anorexia nervosa“ (3 / 1) und „F84.0 Autismus-Spektrum-Störung“ (3 / 1). Mit dieser Auswahl wurden möglichst allgemein bekannte Erkrankungen ausgewählt, für die in früherer Forschung unterschiedlich starke Stigmatisierung festgestellt wurde (Feldmann & Crandall, 2007).

Insgesamt $n = 88$ Psychologiestudierende (74 weiblich, 13 männlich, 1 divers) der Universität des Saarlandes im Alter von 18 bis 30 Jahren ($M = 21.7$, $SD = 2.8$) wurden randomisiert einer der beiden Labeling-Bedingungen zugeordnet (jeweils $n = 44$), in denen hinter der jeweiligen Verhaltensweise in Klammern die Diagnose (oder bei den Pseudo-Items die vermeintliche Diagnose) genannt wurde beziehungsweise keine solche Nennung erfolgte.

Die Dimensionen, auf denen jede Verhaltensweise auf einem Schieberegler von 0-100 bewertet werden sollte, lauteten: „normal – nicht normal“, „selbstverschuldet – nicht selbstverschuldet“,

„selten – häufig“, „behandelbar – nicht behandelbar“, „betrifft mich nicht – betrifft mich“, „nicht behandelbar – behandelbar“, „nicht akzeptabel – akzeptabel“.

Ziel bei der Auswahl dieser Begriffspaare – deren intendierte Bedeutung den Versuchspersonen zu Beginn der Studie erläutert wurde - war unter anderem, typische Inhalte stigmatisierender Einstellungen gegenüber Menschen mit psychischen Erkrankungen zu erfassen. Zum Einen wird den Betroffenen häufig eine eigene Mitschuld an ihrer Erkrankung gegeben (Watson et al., 2003), was mit der Bewertung von Verhalten als „selbstverschuldet“ erfasst werden sollte. Ein weiterer häufiger Bestandteil des Stigmas psychischer Erkrankungen sind herablassende Einstellungen darüber, dass Betroffene nicht eigenständig handeln und entscheiden könnten, sodass ihnen beispielweise der Führerschein oder das politische Wahlrecht entzogen werden sollte (Lauber et al., 2002; Bhugra et al., 2016). Aus diesem Grund wurde mit der Bewertungsdimension „Person ist hilfebedürftig – Person ist selbstständig“ laut Instruktionen an die Versuchspersonen nach der Fähigkeit zu einem „autonome[n], uneingeschränkte[n] Leben“ gefragt.

Wie zu Beginn von Abschnitt 4 dargelegt, führen die inhaltlichen stigmatisierenden Annahmen über Menschen mit psychischen Erkrankungen im Ergebnis zu sozialer Distanz, d. h. einer geringeren Bereitschaft, die Betroffenen im eigenen sozialen Umfeld zu akzeptieren. Diese Einstellungen werden typischerweise mit einer auf Bogardus (1926) zurückgehenden, aus sieben Items bestehenden Skala abgefragt. Eine Nutzung dieser Skala bei jedem der 24 einzelnen Items wurde nicht als praktikabel angesehen. Um das Konstrukt der sozialen Distanz annäherungsweise abzubilden, wurden somit die Dimensionen „nicht akzeptabel“, „nicht normal“ und „betrifft mich nicht“ (im Sinne von: ich kenne dieses Verhalten nicht von mir selbst) aufgenommen.

Pessimistische Ansichten zur Behandelbarkeit einer psychischen Erkrankung spielen eine weniger zentrale Rolle in gängigen Erklärungsmodellen, können jedoch indirekt ebenfalls Annahmen darüber widerspiegeln, wie viel Autonomie man den Betroffenen zuschreibt und sind ein Prädiktor für höhere soziale Distanz (Lebowitz & Ahn, 2012). Schlussendlich wurde die Frage nach der wahrgenommenen Häufigkeit eines Verhaltens mit der Erwartung aufgenommen, dass eine abgrenzende Wirkung diagnostischer Label dazu führen sollte, dass Verhaltensweisen mit Label als ungewöhnlicher oder seltener angesehen werden. Eine Wahrnehmung einer psychischen Erkrankung als selten geht mit einer höheren sozialen Distanz einher (Feldman & Crandall, 2007).

Ergebnisse und Schlussfolgerungen

Zur Überprüfung der Labeling-Effekte wurden gemischte lineare Modelle herangezogen. Neben Alter und Geschlecht als Kontrollvariablen gingen als fixed effects hierbei die Versuchsbedingung

(Label oder kein Label hinter den Symptomen), die Echtheit eines Symptoms (Pseudo-Item oder echtes Diagnosekriterium) und die eigene Betroffenheit der Versuchsperson in Bezug auf diese Diagnose ein. Letztere Variable war mit 1 dummy-kodiert, wenn die Versuchsperson berichtete, von der zu dem jeweiligen Item gehörenden Erkrankung entweder aktuell selbst betroffen zu sein oder in der Vergangenheit betroffen gewesen zu sein (unabhängig von einer tatsächlichen Diagnose). Zudem wurden Interaktionseffekte der beiden letztgenannten Variablen mit der Versuchsbedingung getestet, es wurde somit überprüft, ob sich die An- oder Abwesenheit von Labels je nach Echtheit des Symptoms oder je nach eigener Betroffenheit unterschiedlich auswirkt. Über diese fixed effects hinaus wurden mithilfe von random effects modelliert, inwieweit die Heterogenität der genutzten Items und Diagnosen Varianz in den Bewertungen aufklärt. Dies war bei allen Bewertungskategorien in nicht zu vernachlässigendem Ausmaß der Fall: die gemeinsame Varianzaufklärung betrug rund 42% für die Bewertung von Verhaltensweisen als „normal“, 16% für „selbstverschuldet“, 40% für „häufig“, 8% für „behandelbar“, 28% für „betrifft mich“, 17% für „akzeptabel“ und 28% für „Person ist hilfebedürftig“. Zudem zeigt sich, dass der random effect für die individuellen Versuchspersonen substantiell Varianz aufklärt, besonders auffällig bei den Bewertungskategorien „selbstverschuldet“ mit 34%, „behandelbar“ mit 32% und „akzeptabel“ mit 41%. Diese Hinweise auf starke individuelle Antworttendenzen spiegeln sich auch darin wider, dass die Bewertungen hohe interne Konsistenzen erreichen: Cronbachs Alpha über alle Items hinweg reicht hier von 0.88 bei „betrifft mich nicht – betrifft mich“ bis 0.96 bei „nicht akzeptabel – akzeptabel“.

Eine zusätzliche Modellierung von random slopes - welche abgebildet hätten, wie der Unterschied zwischen den Labeling-Bedingungen sich insbesondere je nach Diagnose unterscheidet – führte mit einer Varianzaufklärung nahe an 0 in allen Fällen zu keinem signifikant verbesserten Modellfit. Somit sind die hier gefundenen Labeling-Effekte als ganz überwiegend unabhängig von der jeweils abgefragten Symptomatik zu bewerten.

Tabelle 5 zeigt die fixed effects je nach Bewertungskriterium. Diese zeigen Labeling-Effekte für drei der sieben Rating-Dimensionen: bei Anwesenheit eines Labels wurden die genannten Verhaltensweisen im Mittel als weniger selbstverschuldet, seltener und weniger typisch für die eigene Person („betrifft mich nicht“) bewertet. Auf der Skala von 1 bis 100 betrugen diese Unterschiede zwischen 7 (Häufigkeit) und 10 (Selbstverschuldung) Einheiten.

Weiterhin zeigen die Ergebnisse, dass Verhaltensweisen teilweise anders bewertet wurden, wenn sie zu einer Erkrankung gehörten, von der die jeweilige Versuchsperson aktueller oder in der Vergangenheit selbst betroffen war. Neben verschiedenen Haupteffekten – wie dem plausiblen, dass betroffene Personen eher angaben, die Verhaltensweise von sich selbst zu kennen – interagierte die

eigene Betroffenheit in fünf Bewertungskategorien auch mit der Experimentalbedingung, sodass die Labeling-Effekte innerhalb jeweils selbst betroffener Versuchspersonen signifikant anders ausfielen als im Rest der Stichprobe. Inhaltlich beeinflusste die Anwesenheit eines Labels die Bewertung bei selbst betroffenen Versuchspersonen dabei jeweils stärker in Richtung der „positiven“ Bewertung beziehungsweise dem rechts angelegten Bewertungspol (d.h. normal, häufig, behandelbar, selbst betroffen und akzeptabel), als dies bei den übrigen Versuchspersonen der Fall war.

Die Art des Items interagierte dagegen in keinem Fall signifikant mit der Versuchsbedingung. Die Anwesenheit eines Labels wirkte sich also jeweils unabhängig davon, ob es sich um ein echtes Symptom oder ein Pseudo-Symptom handelte, ähnlich aus. Auf der Ebene der Haupteffekte unterschieden sich die Pseudo-Items durchaus in einigen Fällen von den genuinen Items. Die Abwesenheit von Interaktionseffekten zeigt jedoch, dass die tatsächliche diagnostische Validität von Verhaltensweisen wie bei schon Matsunaga und Kitamura (2016) sowie Mickelberg und Kollegen (2024) nicht entscheidend für das Auftreten von Labeling-Effekten zu sein scheint.

Tabelle 5. Fixed effects aus den linearen Modellen mit der Bewertung einer Verhaltensweise als Outcome

Bewertung	Effekt	Geschätztes Regression- gewicht	Standard- fehler	95%- Konfidenzintervall		<i>p</i>
				<i>lower</i>	<i>upper</i>	
nicht normal – normal	Intercept	17.072	12.680	-7.795	41.939	
	Alter	0.859	0.455	-0.033	1.751	0.062
	Geschlecht ^a	2.228	3.096	-3.844	8.300	0.474
	Präsenz eines Labels ^b	-3.254	2.587	-8.327	1.820	0.212
	Pseudo-Item ^c	12.249	5.770	0.933	23.564	0.047
	eigene Betroffenheit ^d	-6.373	1.921	-10.140	-2.605	<0.001
	Präsenz eines Labels * Pseudo- Item	-2.686	1.889	-6.389	1.018	0.155
	Präsenz eines Labels * eigene Betroffenheit	5.762	2.567	0.728	10.796	0.025
selbstverschuldet – nicht selbstverschuldet	Intercept	32.392	14.653	3.656	61.127	
	Alter	1.148	0.612	-0.051	2.348	0.064
	Geschlecht ^a	1.832	4.165	-6.336	10.001	0.661
	Präsenz eines Labels ^b	10.409	3.433	3.677	17.141	0.003
	Pseudo-Item ^c	-10.559	3.862	-18.133	-2.985	0.013
	eigene Betroffenheit ^d	-0.922	1.914	-4.677	2.832	0.630

selten - häufig	Präsenz eines Labels * Pseudo-Item	2.752	1.866	-0.907	6.410	0.140
	Präsenz eines Labels * eigene Betroffenheit	1.605	2.557	-3.409	6.620	0.530
	Intercept	55.175	13.127	29.431	80.918	
	Alter	0.214	0.474	-0.716	1.143	0.653
	Geschlecht ^a	-3.158	3.228	-9.490	3.173	0.331
	Präsenz eines Labels ^b	-7.236	2.687	-12.506	-1.966	0.008
	Pseudo-Item ^c	8.417	3.450	1.651	15.184	0.024
	eigene Betroffenheit ^d	-2.197	1.819	-5.764	1.369	0.227
	Präsenz eines Labels * Pseudo-Item	-2.561	1.748	-5.989	0.868	0.143
	Präsenz eines Labels * eigene Betroffenheit	5.387	2.405	0.671	10.103	0.025
	Intercept	82.190	10.555	61.491	102.890	
	Alter	-0.210	0.444	-1.081	0.660	0.637
	Geschlecht ^a	-2.232	3.022	-8.159	3.696	0.462
	Präsenz eines Labels ^b	1.449	2.503	-3.459	6.356	0.564
nicht behandelbar - behandelbar	Pseudo-Item ^c	-2.575	1.220	-4.968	-0.183	0.040
	eigene Betroffenheit ^d	-2.828	1.570	-5.907	0.252	0.072
	Präsenz eines Labels * Pseudo-Item	1.066	1.537	-1.948	4.080	0.488
	Präsenz eines Labels * eigene Betroffenheit	4.260	2.100	0.141	8.378	0.043
	Intercept	54.379	13.035	28.816	79.942	
	Alter	-0.600	0.496	-1.571	0.372	0.230
	Geschlecht ^a	-6.681	3.373	-13.294	-0.067	0.051
	Präsenz eines Labels ^b	-8.368	2.843	-13.944	-2.792	0.004
	Pseudo-Item ^c	2.706	5.423	-7.930	13.341	0.623
	eigene Betroffenheit ^d	10.272	2.382	5.602	14.943	<0.001
	Präsenz eines Labels * Pseudo-Item	-0.040	2.360	-4.669	4.588	0.986
	Präsenz eines Labels * eigene Betroffenheit	7.510	3.189	1.256	13.764	0.019
	Intercept	51.242	16.829	18.238	84.246	
	Alter	0.604	0.717	-0.801	2.009	0.402
nicht akzeptabel - akzeptabel	Geschlecht ^a	-0.163	4.880	-9.734	9.408	0.973
	Präsenz eines Labels ^b	2.255	4.003	-5.594	10.104	0.575

	Pseudo-Item ^c	-10.188	5.699	-21.365	0.989	0.087
	eigene Betroffenheit ^d	-3.853	1.898	-7.575	-0.131	0.043
	Präsenz eines Labels * Pseudo-Item	2.307	1.834	-1.290	5.903	0.209
	Präsenz eines Labels * eigene Betroffenheit	7.306	2.530	2.344	12.268	0.004
Person ist hilfebedürftig – Person ist selbstständig	Intercept	31.280	14.830	2.197	60.363	
	Alter	0.813	0.584	-0.331	1.958	0.167
	Geschlecht ^a	-5.386	3.975	-13.182	2.410	0.179
	Präsenz eines Labels ^b	-6.248	3.288	-12.696	0.200	0.061
	Pseudo-Item ^c	14.546	4.089	6.527	22.565	0.002
	eigene Betroffenheit ^d	-4.203	2.024	-8.171	-0.234	0.038
	Präsenz eines Labels * Pseudo-Item	-3.788	1.966	-7.644	0.067	0.054
	Präsenz eines Labels * eigene Betroffenheit	3.350	2.695	-1.936	8.635	0.214

Anmerkungen. Wertebereich der Bewertungen: 0 – 100. ^a1 = weiblich, 2 = männlich. ^b0 = Präsentation des Items ohne Label, 1 = Präsentation mit Label. ^c0 = Item bildet tatsächliches Diagnosekriterium ab, 1 = Pseudo-Item (siehe Methodenteil). ^d1 = Versuchsperson ist oder war selbst von der Diagnose betroffen (Selbstbericht), 0 = Versuchsperson berichtete keine solche eigene Betroffenheit. SE = Standardfehler, CI = Konfidenzintervall. Signifikante Effekte mit $p < 0.05$ sind hervorgehoben.

Tabelle 6. Modellfitstatistiken zu konfirmatorischen Faktorenanalysen getrennt nach Bewertungskategorie und Experimentalgruppe

Bewertungskategorie	Gruppe	CFI	RMSEA	BIC
nicht normal - normal	mit Label	0.64	0.13	9170.65
	ohne Label	0.54	0.14	9384.57
selbstverschuldet – nicht selbstverschuldet	mit Label	0.76	0.15	9018.99
	ohne Label	0.61	0.15	8976.43
selten - häufig	mit Label	0.70	0.12	8713.77
	ohne Label	0.70	0.13	8254.90
nicht behandelbar - behandelbar	mit Label	0.66	0.16	8277.30
	ohne Label	0.55	0.16	8763.68
betrifft mich – betrifft mich nicht	mit Label	0.71	0.12	9271.79
	ohne Label	0.61	0.12	9779.04
nicht akzeptabel - akzeptabel	mit Label	0.78	0.14	9213.56
	ohne Label	0.73	0.16	8880.06
Person ist hilfebedürftig – Person ist selbstständig	mit Label	0.75	0.13	8951.91
	ohne Label	0.73	0.12	9254.76

Anmerkungen. CFI = Comparative Fit Index, RMSEA = Root Mean Squared Error of Approximation, BIC = Bayesian Information Criterion.

Konfirmatorische Faktorenanalysen (Tabelle 6) deuten einen pauschalisierenden Effekt der diagnostischen Label an: Die Fitstatistiken, insbesondere der CFI, legen überwiegend nahe, dass die Ratings der verschiedenen Items sich stärker entlang ihrer jeweiligen Diagnose zu Faktoren

anordnen, wenn diese Zugehörigkeit in Form des Labels explizit gemacht wird. Zwar ist eine belastbare Faktorenanalyse mit der gegebenen Stichprobengröße kaum als realistisch einzustufen, wie auch der augenscheinlich generell klar hinter den gängigen Grenzwerten zurückbleibende Modellfit zeigt; der Vergleich zwischen den Experimentalbedingungen kann dennoch aussagekräftig bleiben. Allerdings ist der bessere Fit bei Anwesenheit eines Labels nicht durchgängig in allen Fitstatistiken vorhanden, sodass die Schlussfolgerung eines pauschalisierenden Effekts der Label nicht uneingeschränkt gilt.

Die Ergebnisse stützen insgesamt teilweise die Annahme, dass die Nennung eines diagnostischen Labels einen Einfluss auf die Bewertung konkreter Verhaltensweise im Rahmen psychischer Erkrankungen hat. Auch wenn diese These in Form der *modified labeling theory* (Link et al., 1989) bereits seit Jahrzehnten vertreten wird, kann die hier beschriebene Studie III angesichts der uneindeutigen metaanalytischen Ergebnisse bei O'Connor und Kollegen (2021) als wichtige neue Evidenz dafür gesehen werden. Dies gilt insbesondere, da die Labeling-Effekte in Studie III sich unter einigen bisher wenig oder gar nicht untersuchten Voraussetzungen zeigen, im Einzelnen: auf kontextfreier Itemebene statt in Fallvignetten, über diverse Diagnosen hinweg, in verschiedenen Bewertungsdimensionen sowie bei Pseudo-Symptomen ebenso wie bei echten Diagnosekriterien. Die Bandbreite an untersuchten Bewertungsdimensionen liefert eine mögliche Teilerklärung für die heterogene Befundlage in der Literatur zu Labeling-Effekten: in einigen Ratingkategorien zeigten sich keine Effekte, insbesondere auch bei der Frage danach, wie „akzeptabel“ Verhaltensweisen waren. Unter den sieben genutzten Dimensionen in Studie III bildet diese am ehesten unmittelbar Stigmatisierung ab. Dass sich hier also *kein* Effekt der Anwesenheit eines Labels zeigt, verdeutlicht, dass Labeling-Effekte nicht zwingend nur unter dem Blickwinkel betrachtet werden sollten, ob die Nennung einer Diagnose den Betroffenen im Sinne einer Ausgrenzung und Pathologisierung schadet, also der ursprünglichen Stoßrichtung bei Scheff (1966). Im Gegenteil könnte einer der hier gefundenen Effekte sogar als positiv aufgefasst werden, nämlich die deutlich geringere Beurteilung von Verhaltensweisen mit Label als selbstverschuldet. Die Nennung einer Diagnose scheint hier das Verständnis für das Verhalten zu erhöhen, vergleichbar mit den Ergebnissen bei Martinez und Kollegen (2011).

Allerdings könnten diese spezifischen Ergebnisse auch in der hier genutzten Stichprobe begründet sein. Psychologiestudierende können im Hinblick auf Einstellungen gegenüber psychischen Erkrankungen kaum einen typischen Querschnitt der Bevölkerung darstellen. Gleichzeitig ist eine Untersuchung dieses Themas in einer solchen Stichprobe nicht unüblich, da stigmatisierende Ansichten gerade auch in psychosozialen Berufen keine Seltenheit sind (Lauber et al., 2006; Vistorte et al., 2018; Reavley et al., 2014; Valery & Proteau, 2020; Jauch et al., 2023) und auch im

Rahmen von Interventionen oft nicht nachhaltig abgebaut werden können (Lien et al., 2021; Stubbs, 2014). Zudem sind auch Personengruppen mit positiven expliziten Einstellungen nicht vor implizitem Stigma gefeit (Kopera et al., 2015). Interpretiert man Labeling-Effekte als recht automatisch und unbewusst ablaufende Framing-Effekte, wäre es plausibel, sie auf dieser impliziten Ebene zu verorten und damit auch bei explizit positiv gegenüber Betroffenen eingestellten Personen für erwartbar zu halten. Ob sich dies tatsächlich so verhält, muss zukünftige Forschung klären, denn auch wenn die studentische Stichprobe aus diesen Gründen eine Zielgruppe von Interesse darstellt, schränkt sie notwendigerweise die Verallgemeinerbarkeit der Ergebnisse auf die Gesamtbevölkerung ein.

Als praktische Schlussfolgerung für Betroffene und Personen in ihrem Umfeld deuten die gezeigten Ergebnisse darauf hin, dass mit pauschalen Empfehlungen hinsichtlich der Nutzung von diagnostischen Labels vorsichtig umzugehen ist. Die gemischten Ergebnisse aus Studie III lassen vermuten, dass das empfehlenswerte Vorgehen unter anderem davon abhängen könnte, welche Art von Wahrnehmung man erreichen möchte. Eine Darstellung der eigenen Erkrankung etwa als selten oder häufig, als ungewöhnlich oder alltagsnah kann ein zweiseitiges Schwert für Betroffene sein (Gergel, 2014). Beispielsweise ist die wahrgenommene Seltenheit einer psychischen Erkrankung ein substantieller Prädiktor für ihre Stigmatisierung, gleichzeitig kann eine zu starke Normalisierung jedoch auch in unerwünschte Trivialisierung umschlagen (Pavelko & Myrick, 2020). Vor diesem Hintergrund kann der Einfluss eines Labels auf eine Bewertungskategorie wie „selten – häufig“ in Studie III nur schwierig eindeutig als wünschenswert oder nicht wünschenswert eingeordnet werden. Weitere Forschung ist nötig, damit diese schwierigen Abwägungen für Betroffene in Zukunft stärker von empirisch abgeleiteten Handlungsempfehlungen gestützt werden können.

5 Zusammenfassung

Die vorliegende Dissertation untersuchte in drei verschiedenen inhaltlichen Domänen Aspekte der Lebensqualität, wobei in zwei Studien das TwinLife-Panel die Datengrundlage bildete.

In Studie I wurde der Vergleich zwischen begabten Schülern, deren Schulnoten sie als Underachiever definieren, und erfolgreichen begabten Schülern um den Quervergleich zu weiteren nicht erfolgreichen, aber auch nicht hochbegabten Schülern erweitert. Dies ermöglichte eine Ausdifferenzierung der betrachteten Variablen in solche, die in erster Linie mit Schulerfolg im Allgemeinen korrelieren, und solche, die darüber hinaus spezifischer Underachievement als an den kognitiven Fähigkeiten relativierten Misserfolg kennzeichneten. Erstere Unterscheidung war von in der Forschung bereits gut etablierten Korrelaten des Schulerfolgs gekennzeichnet, darunter einem hohen akademischen Selbstkonzept, hoher Lernmotivation und auch hoher Gewissenhaftigkeit. Wurden Underachiever darüber hinaus auch von anderen Niedrigleistern abgegrenzt, waren diese Konstrukte jedoch weniger prominent, wenn auch nicht irrelevant. Dagegen war an dieser Stelle eine durchschnittlich niedrigere Extraversion besonders kennzeichnend für die Gruppe der Underachiever. Neben diesem Vergleich „in beide Richtungen“, der in der bisherigen Forschung größtenteils fehlte, war das Vorgehen in Studie I auch durch eine vergleichsweise restriktive Fallselektion gekennzeichnet, bei der nur rund 3% der Gesamtstichprobe als Underachiever identifiziert wurden. Damit wurde das Risiko fälschlicher Zuordnungen verringert (siehe Ziegler et al., 2012), jedoch auch die Stichprobe so sehr verkleinert, dass die Ergebnisse als wenig gesichert einzuordnen sind, zumal sich Hinweise auf Overfitting fanden.

In Studie II wurde die querschnittlich in verschiedenen früheren Studien gezeigte Korrelation zwischen dem Besuch von Sportvereinen und einer höheren Lebenszufriedenheit erstmals in einem Twin-Difference-Design überprüft, das für mögliche genetische und geteilt umweltbedingte Konfundierungen kontrolliert. Dabei zeigte sich, dass die domänenspezifische Zufriedenheit mit der eigenen Gesundheit und der eigenen Freizeitgestaltung auch in diesem Ansatz mit dem regelmäßigen Besuch von Sportvereinen korreliert blieb. Dies galt jedoch nicht für die allgemeine Lebenszufriedenheit sowie die Zufriedenheit mit dem eigenen Freundeskreis; darüber hinaus blieben auch die genannten Ergebnisse längsschnittlich nicht robust, wenn auch in einer kleineren Stichprobe. Auch hier sind Replikationen wünschenswert, um die, sofern tatsächlich vorhandenen, kleinen Effekte zu untermauern. Dennoch deuten die Ergebnisse von Studie II darauf hin, dass der Besuch von Sportvereinen ein kleiner konkreter Teil des „missing environments“ im Hinblick auf Lebenszufriedenheit sein könnten, d.h. der oft erfolglosen Suche nach konkreten nichtgeteilten Umwelteinflüssen auf Outcomes wie das subjektive Wohlbefinden.

Zuletzt wurde in Studie III ein neuartiger Ansatz zur experimentellen Untersuchung sogenannter Labeling-Effekte im Kontext psychischer Erkrankungen vorgestellt, der die Bewertung einzelner Verhaltensweisen nutzte, um den Einfluss der Nennung diagnostischer Label zu untersuchen. Dabei wurden Labeling-Effekte statt anhand der gängigen Fallvignetten auf der Ebene einzelner Verhaltensweisen untersucht und zudem auf verschiedene Bewertungsdimensionen bezogen. Es zeigte sich ein gemischtes Bild: die Anwesenheit eines Labels ging mit Bewertungen von Verhaltensweisen als seltener, als weniger typisch für die Versuchsperson selbst und als weniger selbstverschuldet einher. Andere Ratings waren zwischen den Experimentalbedingungen dagegen nicht signifikant verschieden, insbesondere die Bewertung eines Verhaltens als normal oder akzeptabel. Diese Ergebnisse waren größtenteils unabhängig von weiteren Randbedingungen. Diese differenzierten Erkenntnisse zeigen, dass die genauen Wirkmechanismen von Labeln und somit auch die zu erwartenden Folgen eines „Outings“ als psychisch erkrankte Person stark vom inhaltlichen Kontext abhängen können.

Die Fragestellungen der einzelnen hier vorgestellten Studien sind ohne Zweifel auch über das Konzept „(subjektive) Lebensqualität“ nur lose thematisch verbunden. Dies ist einerseits aufgrund der Breite des Konstrukts an sich nicht verwunderlich, da die Frage danach, welche Faktoren einen glücklichen und erfolgreichen Lebensweg ausmachen, sehr heterogene Lebensbereiche umfasst (siehe Skevington & Böhnke, 2018). Somit ist auch eine Panelstudie wie TwinLife als Basis zweier der Studien inhaltlich sehr breit gefächert. Über diesen weit gefassten gemeinsamen Nenner hinaus war die Konzeption und Umsetzung der in der vorliegenden Abhandlung enthaltenen Studien stark davon geprägt, mit welchen methodischen Ansätzen die bisherigen Forschungserkenntnissen sinnig ergänzt und erweitert werden könnten. In Studie I betraf dies die valide Identifizierung von Underachievern und die Eingrenzung möglicher Korrelate auf Underachievement als solches; in Studie II die Nutzung des Twin-Difference-Ansatzes als für Konfundierungen kontrollierende Methode; in Studie III die Erforschung von Labeling-Effekten auf Item- statt auf Vignetten-Ebene, zudem ein neuartiges Ergebnis in der Form von Faktorenanalysen, die eine pauschalisierende Wirkweise von Labeln andeuten. In allen Fällen bestätigen die Ergebnisse der vorgestellten Studien teilweise die Erkenntnisse aus der bisherigen Forschungsliteratur, fügen diesen jedoch auch neue Aspekte hinzu, die frühere Annahmen in Frage stellen. So lässt Studie I vermuten, dass neben den klassischen motivationalen Aspekten auch weitere Variablen relevant sein können, wenn eine bessere (Früh-)Entdeckung von Schülern erreicht werden soll, die in der Schule unter ihren kognitiven Möglichkeiten bleiben könnten. Studie II kann die in der Gesamtbevölkerung gut belegten korrelativen Zusammenhänge zwischen sportlicher Aktivität und einer höheren Lebenszufriedenheit teilweise replizieren, gerade für die häufig betrachtete globale

Lebenszufriedenheit findet sich jedoch nur ein durch genetische Konfundierung erklärbarer Effekt. Studie III deutet darauf hin, dass Labeling-Effekte bei psychischen Erkrankungen von der Fragemodalität abhängen könnten und zudem inhaltlich differenzierter betrachtet werden müssen als bisher meist geschehen.

Insgesamt kann die vorliegende Abhandlung damit den Wert neuer Ansätze auch bei bereits stark erforschten Themen wie der subjektiven Lebensqualität und ihren einzelnen Aspekten aufzeigen; viele Fragen in diesem breiten und komplexen Feld sind erkennbar noch nicht abschließend geklärt. Oft betrifft dies die auch in anderen Bereichen der Psychologie als „heiliger Gral“ (siehe Vitaro et al., 2009) zu sehende Frage nach der Kausalität oder nach den genauen Wirkmechanismen hinter gefundenen Zusammenhängen, in den vorgestellten Studien etwa in Bezug auf die Auswirkungen von Sport auf die Lebenszufriedenheit. Gerade verhaltensgenetische Methoden, etwa die in Abschnitt 3 angesprochenen molekulargenetischen Ansätze, könnten hier auch in Zukunft einen wertvollen Beitrag leisten.

Literaturverzeichnis

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). <https://doi.org/10.1176/appi.books.9780890425596>
- Balduf, M. (2009). Underachievement among college students. *Journal of Advanced Academics*, 20(2), 274–294. <https://doi.org/10.1177/1932202X0902000204>
- Barclay, N. L., Eley, T. C., Buysse, D. J., Maughan, B., & Gregory, A. M. (2012). Nonshared environmental influences on sleep quality: A study of monozygotic twin differences. *Behavior Genetics*, 42, 234–244. <https://doi.org/10.1007/s10519-011-9510-1>
- Barry, T. D., Lyman, R. D., & Klinger, L. G. (2002). Academic underachievement and attention-deficit/hyperactivity disorder: The negative impact of symptom severity on school performance. *Journal of School Psychology*, 40(3), 259–283. [https://doi.org/10.1016/S0022-4405\(02\)00100-0](https://doi.org/10.1016/S0022-4405(02)00100-0)
- Bartels, M. (2015). Genetics of wellbeing and its components satisfaction with life, happiness, and quality of life: A review and meta-analysis of heritability studies. *Behavior Genetics*, 45, 137–156. <https://doi.org/10.1007/s10519-015-9713-y>
- Bates, T. C., Maes, H., & Neale, M. C. (2019). umx: Twin and path-based structural equation modeling in R. *Twin Research and Human Genetics*, 22(1), 27–41. <https://doi.org/10.1017/thg.2019.2>
- Becchetti, L., Giachin Ricca, E., & Pelloni, A. (2012). The relationship between social leisure and life satisfaction: Causality and policy implications. *Social Indicators Research*, 108, 453–490. <https://doi.org/10.1007/s11205-011-9887-5>
- Beierlein, C., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2012). *Ein Messinstrument zur Erfassung subjektiver Kompetenzerwartungen: Allgemeine Selbstwirksamkeit Kurzskala (ASKU)* (GESIS-Working Papers, 2012/17). GESIS.
- Bhugra, D., Pathare, S., Gosavi, C., Ventriglio, A., Torales, J., Castaldelli-Maia, J., Gazbar, A. A., Mishara, B., Ng, R., Kandasamy, N., & Robles, R. (2016). Mental illness and the right to vote: A review of legislation across the world. *International Review of Psychiatry*, 28(4), 395–399. <https://doi.org/10.1080/09540261.2016.1211096>
- Bogardus, E. S. (1926). Social distance in the city. *Proceedings and Publications of the American Sociological Society*, 20, 40–46.
- Breuer, C., Feiler, S., & Wicker, P. (2015). Sport clubs in Germany. In C. Breuer, R. Hoekman, S.

- Nagel, & H. van der Werff (Eds.), *Sport clubs in Europe: A cross-national comparative perspective* (pp. 187–208). Springer. https://doi.org/10.1007/978-3-319-17635-2_11
- Buecker, S., Simacek, T., Ingwersen, B., Terwiel, S., & Simonsmeier, B. A. (2021). Physical activity and subjective well-being in healthy individuals: A meta-analytic review. *Health Psychology Review, 15*(4), 574–592. <https://doi.org/10.1080/17437199.2020.1760728>
- Camfield, L., & Skevington, S. M. (2008). On subjective well-being and quality of life. *Journal of health psychology, 13*(6), 764–775. <https://doi.org/10.1177/1359105308093860>
- Carlin, J. B., Gurrin, L. C., Sterne, J. A., Morley, R., & Dwyer, T. (2005). Regression models for twin studies: A critical review. *International Journal of Epidemiology, 34*(5), 1089–1099. <https://doi.org/10.1093/ije/dyi153>
- Carlsson, S., Andersson, T., Lichtenstein, P., Michaëlsson, K., & Ahlbom, A. (2006). Genetic effects on physical activity: Results from the Swedish Twin Registry. *Medicine & Science in Sports & Exercise, 38*(8), 1396–1401. <https://doi.org/10.1249/01.mss.0000228941.17034.c1>
- Castejón, J. L., Gilar, R., Veas, A., & Miñano, P. (2016). Differences in learning strategies, goal orientations, and self-concept between overachieving, normal-achieving, and underachieving secondary students. *Frontiers in Psychology, 7*, Article 1438. <https://doi.org/10.3389/fpsyg.2016.01438>
- Charlemagne-Badal, S. J., Lee, J. W., Butler, T. L., & Fraser, G. E. (2015). Conceptual domains included in wellbeing and life satisfaction instruments: A review. *Applied Research in Quality of Life, 10*(2), 305–328. <https://doi.org/10.1007/s11482-014-9306-6>
- Colangelo, N., Kerr, B., Christensen, P., & Maxey, J. (1993). A comparison of gifted underachievers and gifted high achievers. *Gifted Child Quarterly, 37*(4), 155–160. <https://doi.org/10.1177/001698629303700404>
- Cone, T. E., & Wilson, L. R. (1981). Quantifying a severe discrepancy: A critical analysis. *Learning Disability Quarterly, 4*(4), 359–371. <https://doi.org/10.2307/1510688>
- Costa, P. T., Jr., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13*(6), 653–665. [https://doi.org/10.1016/0191-8869\(92\)90236-I](https://doi.org/10.1016/0191-8869(92)90236-I)
- Cummins, R. A. (2000). Objective and subjective quality of life: An interactive model. *Social indicators research, 52*(1), 55–72. <https://doi.org/10.1023/A:1007027822521>
- Dickhäuser, O., Schöne, C., Spinath, B., & Stiensmeier-Pelster, J. (2002). Die Skalen zum

akademischen Selbstkonzept: Konstruktion und Überprüfung eines neuen Instrumentes. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 23(4), 393–405.
<https://doi.org/10.1024//0170-1789.23.4.393>

- Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13
- Diener, E., Lucas, R. E., & Oishi, S. (2018). Advances and open questions in the science of subjective well-being. *Collabra: Psychology*, 4(1), Article 15.
<https://doi.org/10.1525/collabra.115>
- Dings, A., & Kubillus, L. K. (in press). Beyond case vignettes: Do diagnostic labels affect how symptoms of mental illness are perceived? *Journal of Mental Health Disorders*, 5(1), 1–14.
<https://doi.org/10.33696/mentalhealth.5.030>
- Dings, A., & Spinath, F. M. (2021). Motivational and personality variables distinguish academic underachievers from high achievers, low achievers, and overachievers. *Social Psychology of Education*, 24(6), 1461–1485. <https://doi.org/10.1007/s11218-021-09659-2>
- Dings, A., & Spinath, F. M. (2024). Sports club participation impacts life satisfaction in adolescence: A twin study. *Psychology of Sport and Exercise*, 73, Article 102639.
<https://doi.org/10.1016/j.psychsport.2024.102639>
- Edgerton, J.D., Roberts, L.W., von Below, S. (2012). Education and Quality of Life. In: Land, K., Michalos, A., Sirgy, M. (eds) *Handbook of Social Indicators and Quality of Life Research*. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-2421-1_12.
- Eime, R. M., Harvey, J. T., Brown, W. J., & Payne, W. R. (2010). Does sports club participation contribute to health-related quality of life? *Medicine & Science in Sports & Exercise*, 42(5), 1022–1028. <https://doi.org/10.1249/MSS.0b013e3181c3adaa>
- Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1), 51–76. <https://doi.org/10.1111/j.1469-1809.1965.tb00500.x>
- Feldman, D. B., & Crandall, C. S. (2007). Dimensions of mental illness stigma: What about mental illness causes social rejection? *Journal of Social and Clinical Psychology*, 26(2), 137–154.
<https://doi.org/10.1521/jscp.2007.26.2.137>
- Figg, S. D., Rogers, K. B., McCormick, J., & Low, R. (2012). Differentiating low performance of the gifted learner: Achieving, underachieving, and selective consuming students. *Journal of*

- Flage, A., & Le Gallo, J. (2025). Discrimination against people with mental, physical or visual disabilities in the French rental housing market: field experiment. *Housing Studies*, 40(1), 116–138. <https://doi.org/10.1080/02673037.2023.2266412>
- Forbes, M. K., Neo, B., Nezami, O. M., Fried, E. I., Faure, K., Michelsen, B., Twose, C., Dras, M., Batterham, P. J., Baillie, A. J., Krueger, R. F., & Christensen, H. (2024). Elemental psychopathology: Distilling constituent symptoms and patterns of repetition in the diagnostic criteria of the DSM-5. *Psychological Medicine*, 54(5), 886–894. <https://doi.org/10.1017/S0033291723002969>
- Fox, A. B., Earnshaw, V. A., Taverna, E. C., & Vogt, D. (2018). Conceptualizing and measuring mental illness stigma: The mental illness stigma framework and critical review of measures. *Stigma and Health*, 3(4), 348–376. <https://doi.org/10.1037/sah0000104>
- Fox, A. B., Vogt, D., Boyd, J. E., Earnshaw, V. A., Janio, E. A., Davis, K., Broussard, K. A., Taverna, E. C., & Sorkin, D. H. (2021). Mental illness, problem, disorder, distress: Does terminology matter when measuring stigma? *Stigma and Health*, 6(4), 419–430. <https://doi.org/10.1037/sah0000310>
- Gergel, T. L. (2014). Too similar, too different: The paradoxical dualism of psychiatric stigma. *BJPsych Bulletin*, 38(4), 148–151. <https://doi.org/10.1192/pb.bp.113.044693>
- Gerlitz, J. Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. *DIW Research Notes*, 4, 1–44. <https://doi.org/10.1016/j.paid.2005.01.016>
- Godfrey, E., & Koutsouris, G. (2024). Is personality overlooked in educational psychology? Educational experiences of secondary-school students with introverted personality styles. *Educational Psychology in Practice*, 40(2), 159–184. <https://doi.org/10.1080/02667363.2023.2287524>
- Görzig, A., & Ryan, L. N. (2022). The different faces of mental illness stigma: Systematic variation of stereotypes, prejudice and discrimination by type of illness. *Scandinavian Journal of Psychology*, 63(5), 545–554. <https://doi.org/10.1111/sjop.12827>
- Gottschling, J. (2017). *Documentation TwinLife data: Cognitive abilities* (TwinLife Technical Report Series). <https://pub.uni-bielefeld.de/record/2913820>
- Gove, W. R. (1979). The labeling versus the psychiatric explanation of mental illness: A debate that has become substantively irrelevant: Reply. *Journal of Health and Social Behavior*, 20(3), 301–

304. <https://doi.org/10.1080/02667363.2023.2287524>

- Granello, D. H., & Gibbs, T. A. (2016). The power of language and labels: "The mentally ill" versus "people with mental illnesses." *Journal of Counseling & Development, 94*(1), 31–40. <https://doi.org/10.1002/jcad.12059>
- Gringras, P., & Chen, W. (2001). Mechanisms for differences in monozygous twins. *Early Human Development, 64*(2), 105–117. [https://doi.org/10.1016/S0378-3782\(01\)00171-2](https://doi.org/10.1016/S0378-3782(01)00171-2)
- Hahn, E., Gottschling, J., Bleidorn, W., Kandler, C., Spengler, M., Kornadt, A. E., Schulz, S., Schunck, R., Baier, T., Krell, K., Zimmermann, J., Franz, M., Lang, V., Lenau, F., Mund, M., Hufer, A., Mönkediek, B., Weigel, L., & Spinath, F. M. (2016). What drives the development of social inequality over the life course? The German TwinLife study. *Twin Research and Human Genetics, 19*(6), 659–672. <https://doi.org/10.1017/thg.2016.76>
- Hahn, E., Johnson, W., & Spinath, F. M. (2013). Beyond the heritability of life satisfaction—The roles of personality and twin-specific influences. *Journal of Research in Personality, 47*(6), 757–767. <https://doi.org/10.1016/j.jrp.2013.07.003>
- Hanses, P., & Rost, D. H. (1996). *Das "Drama der hochbegabten Underachiever:" "gewöhnliche" oder "aussergewöhnliche" Underachiever?* Fachbereich Psychologie der Philipps-Universität.
- Harden, K. P. (2021). "Reports of my death were greatly exaggerated": Behavior genetics in the postgenomic era. *Annual Review of Psychology, 72*(1), 37–60. <https://doi.org/10.1146/annurev-psych-052220-103822>
- Hastuti, R., & Timming, A. R. (2021). An inter-disciplinary review of the literature on mental illness disclosure in the workplace: Implications for human resource management. *The International Journal of Human Resource Management, 32*(15), 3302–3338. <https://doi.org/10.1080/09585192.2021.1875494>
- Hatemi, P. K., Medland, S. E., Klemmensen, R., Oskarsson, S., Littvay, L., Dawes, C. T., Verhulst, B., McDermott, R., Sonne, C., Jørgensen, K. M., Krauss, S., Hoffman, E., Bäck, H., Farde, M., Nilsson, L., Strandberg, K., Morley, K. I., Zhu, G., Suhay, E., ... Martin, N. G. (2014). Genetic influences on political ideologies: Twin analyses of 19 measures of political ideologies from five democracies and genome-wide findings from three populations. *Behavior Genetics, 44*(3), 282–294. <https://doi.org/10.1007/s10519-014-9648-8>
- Hattie, J., & Yates, G. C. (2013). *Visible learning and the science of how we learn*. Routledge. <https://doi.org/10.4324/9781315885025>

- Hoover-Schultz, B. (2005). Gifted underachievement: Oxymoron or educational enigma? *Gifted Child Today*, 28(2), 46–49. <https://doi.org/10.4219/gct-2005-171>
- Horwitz, A. (1979). Models, muddles, and mental illness labeling. *Journal of Health and Social Behavior*, 20(3), 296–300. <https://doi.org/10.2307/2136454>
- Imhoff, R. (2016). Zeroing in on the effect of the schizophrenia label on stigmatizing attitudes: A large-scale study. *Schizophrenia Bulletin*, 42(2), 456–463. <https://doi.org/10.1093/schbul/sbv137>
- Jackson, R. L., & Jung, J. Y. (2022). The identification of gifted underachievement: Validity evidence for the commonly used methods. *British Journal of Educational Psychology*, 92(3), 1133–1159. <https://doi.org/10.1111/bjep.12492>
- Jauch, M., Occhipinti, S., & O'Donovan, A. (2023). The stigmatization of mental illness by mental health professionals: Scoping review and bibliometric analysis. *PLOS ONE*, 18(1), Article e0280739. <https://doi.org/10.1371/journal.pone.0280739>
- Johnson, W., McGue, M., & Iacono, W. G. (2006). Genetic and environmental influences on academic achievement trajectories during adolescence. *Developmental Psychology*, 42(3), 514–532. <https://doi.org/10.1037/0012-1649.42.3.514>
- Johnson, W., Turkheimer, E., Gottesman, I. I., & Bouchard, T. J., Jr. (2009). Beyond heritability: Twin studies in behavioral research. *Current Directions in Psychological Science*, 18(4), 217–220. <https://doi.org/10.1111/j.1467-8721.2009.01639.x>
- Jones, A. M. (2011). Disclosure of mental illness in the workplace: A literature review. *American Journal of Psychiatric Rehabilitation*, 14(3), 212–229. <https://doi.org/10.1080/15487768.2011.598101>
- Jones, S., & Myhill, D. (2004). Seeing things differently: Teachers' constructions of underachievement. *Gender and Education*, 16(4), 531–546. <http://doi.org/10.1080/09540250042000300411>
- Jorm, A. F., & Oh, E. (2009). Desire for social distance from people with mental disorders. *Australian & New Zealand Journal of Psychiatry*, 43(3), 183–200. <https://doi.org/10.1080/00048670802653349>
- Jorm, A. F., Reavley, N. J., & Ross, A. M. (2012). Belief in the dangerousness of people with mental disorders: a review. *Australian & New Zealand Journal of Psychiatry*, 46(11), 1029–1045. <https://doi.org/10.1177/0004867412442406>

- Kawakami, K., Procopio, F., Rimfeld, K., Malanchini, M., von Stumm, S., Asbury, K., & Plomin, R. (2024). Exploring the genetic prediction of academic underachievement and overachievement. *npj Science of Learning*, 9(1), Article 39. <https://doi.org/10.1038/s41539-024-00251-9>
- Keller, M. C., Medland, S. E., & Duncan, L. E. (2010). Are extended twin family designs worth the trouble? A comparison of the bias, precision, and accuracy of parameters estimated in four twin family models. *Behavior Genetics*, 40(3), 377–393. <https://doi.org/10.1007/s10519-009-9320-x>
- Koenig, L. B., McGue, M., Krueger, R. F., & Bouchard, T. J., Jr. (2005). Genetic and environmental influences on religiousness: Findings for retrospective and current religiousness ratings. *Journal of Personality*, 73(2), 471–488. <https://doi.org/10.1111/j.1467-6494.2005.00316.x>
- Kopera, M., Suszek, H., Bonar, E., Myszk, M., Gmaj, B., Ilgen, M., & Wojnar, M. (2015). Evaluating explicit and implicit stigma of mental illness in mental health professionals and medical students. *Community Mental Health Journal*, 51(5), 628–634. <https://doi.org/10.1007/s10597-014-9796-6>
- Korhonen, T., Kujala, U. M., Rose, R. J., & Kaprio, J. (2009). Physical activity in adolescence as a predictor of alcohol and illicit drug use in early adulthood: A longitudinal population-based twin study. *Twin Research and Human Genetics*, 12(3), 261–268. <https://doi.org/10.1375/twin.12.3.261>
- Kujala, U. M., Kaprio, J., Sarna, S., & Koskenvuo, M. (1998). Relationship of leisure-time physical activity and mortality: The Finnish twin cohort. *JAMA*, 279(6), 440–444. <https://doi.org/10.1001/jama.279.6.440>
- Landis, R. N., & Reschly, A. L. (2013). Reexamining gifted underachievement and dropout through the lens of student engagement. *Journal for the Education of the Gifted*, 36(2), 220–249. <https://doi.org/10.1177/0162353213480864>
- Lang, V., & Kottwitz, A. (2020). The socio-demographic structure of the first wave of the TwinLife panel study: A comparison with the microcensus. *Methods, Data, Analyses*, 14(1), 5–34. <https://doi.org/10.12758/mda.2020.02>
- Lau, K. L., & Chan, D. W. (2001). Identification of underachievers in Hong Kong: Do different methods select different underachievers? *Educational Studies*, 27(2), 187–200. <https://doi.org/10.1080/03055690120050419>
- Lauber, C., Nordt, C., Braunschweig, C., & Rössler, W. (2006). Do mental health professionals

- stigmatize their patients? *Acta Psychiatrica Scandinavica*, 113(Suppl. 429), 51–59.
<https://doi.org/10.1111/j.1600-0447.2005.00718.x>
- Lauber, C., Nordt, C., Falcato, L., & Rössler, W. (2002). Public attitude to compulsory admission of mentally ill people. *Acta Psychiatrica Scandinavica*, 105(5), 385–389.
<https://doi.org/10.1034/j.1600-0447.2002.1o267.x>
- Lebowitz, M. S., & Ahn, W. K. (2012). Combining biomedical accounts of mental disorders with treatability information to reduce mental illness stigma. *Psychiatric Services*, 63(5), 496–499.
<https://doi.org/10.1176/appi.ps.201100265>
- Lent, R. W., Singley, D., Sheu, H. B., Gainor, K. A., Brenner, B. R., Treistman, D., & Ades, L. (2005). Social cognitive predictors of domain and life satisfaction: Exploring the theoretical precursors of subjective well-being. *Journal of Counseling Psychology*, 52(3), 429–442.
<https://doi.org/10.1037/0022-0167.52.3.429>
- Leskinen, T., Sipilä, S., Alen, M., Cheng, S., Pietiläinen, K. H., Usenius, J. P., Suominen, H., Kovanen, V., Kainulainen, H., Kaprio, J., & Kujala, U. M. (2009). Leisure-time physical activity and high-risk fat: A longitudinal population-based twin study. *International Journal of Obesity*, 33(11), 1211–1218. <https://doi.org/10.1038/ijo.2009.170>
- Lien, Y. Y., Lin, H. S., Lien, Y. J., Tsai, C. H., Wu, T. T., Li, H., & Tu, Y. K. (2021). Challenging mental illness stigma in healthcare professionals and students: A systematic review and network meta-analysis. *Psychology & Health*, 36(6), 669–684.
<https://doi.org/10.1080/08870446.2020.1828413>
- Link, B. G., Cullen, F. T., Struening, E., Shrout, P. E., & Dohrenwend, B. P. (1989). A modified labeling theory approach to mental disorders: An empirical assessment. *American Sociological Review*, 54(3), 400–423. <https://doi.org/10.2307/2095613>
- Long, R. F., & Huebner, E. S. (2014). Differential validity of global and domain-specific measures of life satisfaction in the context of schooling. *Child Indicators Research*, 7, 671–694.
<https://doi.org/10.1007/s12187-013-9231-5>
- Maher, B. (2008). The case of the missing heritability. *Nature*, 456(7218), 18–21.
<https://doi.org/10.1038/456018a>
- Maher, J. P., & Conroy, D. E. (2017). Daily life satisfaction in older adults as a function of (in)activity. *The Journals of Gerontology: Series B*, 72(4), 593–602.
<https://doi.org/10.1093/geronb/gbv086>

- Manago, B., & Mize, T. D. (2024). What is the effect of an inferred mental illness label on stigma? Theoretical and empirical challenges. *Socius*, 10, Article 23780231241274237. <https://doi.org/10.1177/23780231241274237>
- Mann, F. D., DeYoung, C. G., Tiberius, V., & Krueger, R. F. (2019). Social-relational exposures and well-being: Using multivariate twin data to rule-out heritable and shared environmental confounds. *Journal of Research in Personality*, 83, Article 103880. <https://doi.org/10.1016/j.jrp.2019.103880>
- Martinez, A. G., Piff, P. K., Mendoza-Denton, R., & Hinshaw, S. P. (2011). The power of a label: Mental illness diagnoses, ascribed humanity, and social rejection. *Journal of Social and Clinical Psychology*, 30(1), 1–23. <https://doi.org/10.1521/jscp.2011.30.1.1>
- Masland, S. R., & Null, K. E. (2022). Effects of diagnostic label construction and gender on stigma about borderline personality disorder. *Stigma and Health*, 7(1), 89–99. <https://doi.org/10.1037/sah0000305>
- Matsunaga, A., & Kitamura, T. (2016). The effects of symptoms, diagnostic labels, and education in psychiatry on the stigmatization towards schizophrenia: A questionnaire survey among a lay population in Japan. *Mental Illness*, 8(1), Article 6502. <https://doi.org/10.4081/mi.2016.6344>
- Mazrekaj, D., De Witte, K., & Triebs, T. P. (2022). Mind the gap: Measuring academic underachievement using stochastic frontier analysis. *Exceptional Children*, 88(4), 442–459. <https://doi.org/10.1177/00144029211073524>
- McAdams, T. A., Rijdsdijk, F. V., Zavos, H. M., & Pingault, J. B. (2021). Twins and causal inference: Leveraging nature's experiment. *Cold Spring Harbor Perspectives in Medicine*, 11(6), Article a039552. <https://doi.org/10.1101/cshperspect.a039552>
- McCall, R. B., Beach, S. R., & Lau, S. (2000). The nature and correlates of underachievement among elementary schoolchildren in Hong Kong. *Child Development*, 71(3), 785–801. <https://doi.org/10.1111/1467-8624.00185>
- McCall, R. B., Evahn, C., & Kratzer, L. (1992). *High school underachievers: What do they achieve as adults?* Sage Publications.
- McCoach, D. B., & Siegle, D. (2003a). Factors that differentiate underachieving gifted students from high-achieving gifted students. *Gifted Child Quarterly*, 47(2), 144–154. <https://doi.org/10.1177/001698620304700205>
- McCoach, D. B., & Siegle, D. (2003b). The school attitude assessment survey-revised: A new

- instrument to identify academically able students who underachieve. *Educational and Psychological Measurement*, 63(3), 414–429. <https://doi.org/10.1177/0013164403063003005>
- McCoach, D. B., Siegle, D., & Rubenstein, L. D. (2020). Pay attention to inattention: Exploring ADHD symptoms in a sample of underachieving gifted students. *Gifted Child Quarterly*, 64(2), 100–116. <https://doi.org/10.1177/0016986219901320>
- McGue, M., & Bouchard, T. J., Jr. (1984). Adjustment of twin data for the effects of age and sex. *Behavior Genetics*, 14(4), 325–343. <https://doi.org/10.1007/BF01080045>
- McGue, M., Osler, M., & Christensen, K. (2010). Causal inference and observational research: The utility of twins. *Perspectives on Psychological Science*, 5(5), 546–556. <https://doi.org/10.1177/1745691610383511>
- Mickelberg, A. J., Walker, B., Ecker, U. K., & Fay, N. (2024). Helpful or harmful? The effect of a diagnostic label and its later retraction on person impressions. *Acta Psychologica*, 248, Article 104420. <https://doi.org/10.1016/j.actpsy.2024.104420>
- Misheva, V. (2016). What determines emotional well-being? The role of adverse experiences: Evidence using twin data. *Journal of Happiness Studies*, 17, 1921–1937. <https://doi.org/10.1007/s10902-015-9678-9>
- Mofield, E. L., & Parker Peters, M. (2022). Understanding the interplay of psychosocial competencies in talent development: Typologies and differences for gifted students. *Roeper Review*, 44(3), 144–156. <https://doi.org/10.1080/02783193.2022.2071368>
- Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, 21(15), 3301–3307. <https://doi.org/10.1093/bioinformatics/bti499>
- Mönkediek, B., Lang, V., Weigel, L., Baum, M. A., Eifler, E. F., Hahn, E., Hufer, A., Krell, K., Lenau, F., Lorenschat, M., Melfsen, S., Müller, R., Pascher, J., Patzelt, A., Peters, A. L., Pötzschke, S., Rammstedt, B., Schacht, D., Schulz, S., ... Spinath, F. M. (2019). The German twin family panel (TwinLife). *Twin Research and Human Genetics*, 22(6), 540–547. <https://doi.org/10.1017/thg.2019.63>
- Mutz, M., Reimers, A. K., & Demetriou, Y. (2021). Leisure time sports activities and life satisfaction: Deeper insights based on a representative survey from Germany. *Applied Research in Quality of Life*, 16, 2155–2171. <https://doi.org/10.1007/s11482-020-09866-7>
- Nes, R. B., Czajkowski, N., Røysamb, E., Reichborn-Kjennerud, T., & Tambs, K. (2008). Well-

- being and ill-being: Shared environments, shared genes? *Journal of Positive Psychology*, 3(4), 253–265. <https://doi.org/10.1080/17439760802399323>
- O'Connor, C., Brassil, M., O'Sullivan, S., Seery, C., & Nearchou, F. (2022). How does diagnostic labelling affect social responses to people with mental illness? A systematic review of experimental studies using vignette-based designs. *Journal of Mental Health*, 31(1), 115–130. <https://doi.org/10.1080/09638237.2021.1922653>
- Pavelko, R. L., & Myrick, J. G. (2020). Measuring trivialization of mental illness: Developing a scale of perceptions that mental illness symptoms are beneficial. *Health Communication*, 35(13), 1589–1600. <https://doi.org/10.1080/10410236.2019.1573296>
- Pavot, W., & Diener, E. (2009). Review of the satisfaction with life scale. In E. Diener (Ed.), *Assessing well-being: The collected works of Ed Diener* (pp. 101–117). Springer. https://doi.org/10.1007/978-90-481-2354-4_5
- Pedišić, Ž., Greblo, Z., Phongsavan, P., Milton, K., & Bauman, A. E. (2015). Are total, intensity- and domain-specific physical activity levels associated with life satisfaction among university students? *PLOS ONE*, 10(2), Article e0118137. <https://doi.org/10.1371/journal.pone.0118137>
- Pescosolido, B. A., Medina, T. R., Martin, J. K., & Long, J. S. (2013). The "backbone" of stigma: Identifying the global core of public prejudice associated with mental illness. *American Journal of Public Health*, 103(5), 853–860. <https://doi.org/10.2105/AJPH.2012.301147>
- Pheister, M., Peters, R. M., & Wrzosek, M. I. (2020). The impact of mental illness disclosure in applying for residency. *Academic Psychiatry*, 44(5), 554–561. <https://doi.org/10.1007/s40596-020-01227-8>
- Pietsch, S., Linder, S., & Jansen, P. (2022). Well-being and its relationship with sports and physical activity of students during the coronavirus pandemic. *German Journal of Exercise and Sport Research*, 52(1), 50–57. <https://doi.org/10.1007/s12662-021-00750-6>
- Plomin, R. (2024). Nonshared environment: Real but random. *JCPP Advances*, 4(3), Article e12229. <https://doi.org/10.1002/jcv2.12229>
- Plomin, R., Asbury, K., & Dunn, J. (2001). Why are children in the same family so different? Nonshared environment a decade later. *The Canadian Journal of Psychiatry*, 46(3), 225–233. <https://doi.org/10.1177/070674370104600302>
- Polderman, T. J., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years

- of twin studies. *Nature Genetics*, 47(7), 702–709. <https://doi.org/10.1038/ng.3285>
- Posthuma, D. (2009). Multivariate genetic analysis. In Y. K. Kim (Ed.), *Handbook of behavior genetics* (pp. 47–59). Springer. https://doi.org/10.1007/978-0-387-76727-7_4
- Preckel, F., Holling, H., & Vock, M. (2006). Academic underachievement: Relationship with cognitive motivation, achievement motivation, and conscientiousness. *Psychology in the Schools*, 43(3), 401–411. <https://doi.org/10.1002/pits.20154>
- Reavley, N. J., Mackinnon, A. J., Morgan, A. J., & Jorm, A. F. (2014). Stigmatising attitudes towards people with mental disorders: A comparison of Australian health professionals with the general community. *Australian & New Zealand Journal of Psychiatry*, 48(5), 433–441. <https://doi.org/10.1177/0004867413500351>
- Reis, S. M., & McCoach, D. B. (2000). The underachievement of gifted students: What do we know and where do we go? *Gifted Child Quarterly*, 44(3), 152–170. <https://doi.org/10.1177/001698620004400302>
- Rijsdijk, F. V., & Sham, P. C. (2002). Analytic approaches to twin data using structural equation models. *Briefings in Bioinformatics*, 3(2), 119–133. <https://doi.org/10.1093/bib/3.2.119>
- Ritchotte, J. A., Matthews, M. S., & Flowers, C. P. (2014). The validity of the achievement-orientation model for gifted middle school students: An exploratory study. *Gifted Child Quarterly*, 58(3), 183–198. <https://doi.org/10.1177/0016986214534890>
- Rost, D. H., & Hanses, P. (1998). Wer nichts leistet, ist nicht begabt? Zur Identifikation hochbegabter Underachiever durch Lehrkräfte [Not achieving—not gifted? About the identification of gifted underachievers by teacher-ratings]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 29(2), 167–177.
- Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, 53, 118–137. <https://doi.org/10.1016/j.intell.2015.09.002>
- Røysamb, E., & Tambs, K. (2016). The beauty, logic and limitations of twin studies. *Norsk Epidemiologi*, 26(1–2), 35–46. <https://doi.org/10.5324/nje.v26i1-2.2014>
- Rubenstein, L. D., Siegle, D., Reis, S. M., McCoach, D. B., & Burton, M. G. (2012). A complex quest: The development and research of underachievement interventions for gifted students. *Psychology in the Schools*, 49(7), 678–694. <https://doi.org/10.1002/pits.21620>

- Rüsch, N., Angermeyer, M. C., & Corrigan, P. W. (2005). Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma. *European Psychiatry*, 20(8), 529–539. <https://doi.org/10.1016/j.eurpsy.2005.04.004>
- Rüsch, N., Heland-Graef, M., & Berg-Peer, J. (2020). *Das Stigma psychischer Erkrankung: Strategien gegen Diskriminierung und Ausgrenzung* [Mental illness stigma: Strategies against discrimination and exclusion]. Elsevier Health Sciences.
- Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual review of psychology*, 52(1), 141–166. <https://doi.org/10.1146/annurev.psych.52.1.141>
- Sadler, M. S., Kaye, K. E., & Vaughn, A. A. (2015). Competence and warmth stereotypes prompt mental illness stigma through emotions. *Journal of Applied Social Psychology*, 45(11), 602–612. <https://doi.org/10.1111/jasp.12316>
- Sapir, E. (1929). The status of linguistics as a science. *Language*, 5(4), 207–214. <https://doi.org/10.2307/409588>
- Scheff, T. (1966). *Being mentally ill: A sociological theory*. Aldine. <https://doi.org/10.4324/9781351328241>
- Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegate, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, 11(9), 1141–1152. <https://doi.org/10.1111/2041-210X.13434>
- Schimmack, U., Diener, E., & Oishi, S. (2002). Life-satisfaction is a momentary judgment and a stable personality characteristic: The use of chronically accessible and stable sources. *Journal of Personality*, 70(3), 345–384. <https://doi.org/10.1111/1467-6494.05006>
- Schleithoff, F. (2015). Noteninflation im deutschen Schulsystem—Macht das Abitur hochschulreif? [Grade inflation in the German school system]. *ORDO*, 66(1), 3–26. <https://doi.org/10.1515/ordo-2015-0102>
- Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, 98, 146–151. <https://doi.org/10.1016/j.jclinepi.2017.12.006>
- Schmiedeberg, C., & Schröder, J. (2017). Leisure activities and life satisfaction: An analysis with German panel data. *Applied Research in Quality of Life*, 12, 137–151. <https://doi.org/10.1007/s11482-016-9458-7>

- Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, 143(6), 565–600.
<https://doi.org/10.1037/bul0000098>
- Schnittker, J. (2008). Happiness and success: Genes, families, and the psychological effects of socioeconomic position and social support. *American Journal of Sociology*, 114(S1), S233–S259. <https://doi.org/10.1086/592424>
- Schnyder, N., Panczak, R., Groth, N., & Schultze-Lutter, F. (2017). Association between mental health-related stigma and active help-seeking: systematic review and meta-analysis. *The British Journal of Psychiatry*, 210(4), 261–268. <https://doi.org/10.1192/bjp.bp.116.189464>
- Schomerus, G., Sander, C., Schindler, S., Baumann, E., & Angermeyer, M. C. (2023). Public attitudes towards protecting the human rights of people with mental illness: A scoping review and data from a population trend study in Germany. *International Review of Psychiatry*, 35(2), 167–179. <https://doi.org/10.1080/09540261.2022.2087494>
- Siegle, D. (2018). Understanding underachievement. In S. I. Pfeiffer, E. Shaunessy-Dedrick, & M. Foley-Nicpon (Eds.), *APA handbook of giftedness and talent* (pp. 285–297). American Psychological Association. <https://doi.org/10.1037/0000038-019>
- Siegle, D., & McCoach, D. B. (2005). Making a difference: Motivating gifted students who are not achieving. *Teaching Exceptional Children*, 38(1), 22–27.
<https://doi.org/10.1177/004005990503800104>
- Siegle, D., DaVia Rubenstein, L., & McCoach, D. B. (2020). Do you know what I'm thinking? A comparison of teacher and parent perspectives of underachieving gifted students' attitudes. *Psychology in the Schools*, 57(10), 1596–1614. <https://doi.org/10.1002/pits.22345>
- Sjölander, A., Frisell, T., & Öberg, S. (2012). Causal interpretation of between-within models for twin research. *Epidemiologic Methods*, 1(1), 217–237. <https://doi.org/10.1515/2161-962X.1015>
- Skevington, S. M., & Böhnke, J. R. (2018). How is subjective well-being related to quality of life? Do we need two concepts and both measures? *Social science & medicine*, 206, 22–30.
<https://doi.org/10.1016/j.socscimed.2018.04.005>
- Smith, E. (2007). Considering the experiences of 'underachieving' and 'overachieving' students. *International Journal of Research & Method in Education*, 30(1), 19–32.
<https://doi.org/10.1080/17437270701207702>
- Smith, G. D. (2011). Epidemiology, epigenetics and the 'Gloomy Prospect': Embracing randomness

in population health research and practice. *International Journal of Epidemiology*, 40(3), 537–562. <https://doi.org/10.1093/ije/dyr117>

Sparfeldt, J. R., Buch, S. R., & Rost, D. H. (2010). Klassenprimus bei durchschnittlicher Intelligenz—Overachiever auf dem Gymnasium [Excellent scholastic achievement with average intelligence—Overachievement in secondary school]. *Zeitschrift für Pädagogische Psychologie*, 24(2), 147–155. <https://doi.org/10.1024/1010-0652/a000012>

Spengler, S., & Woll, A. (2013). The more physically active, the healthier? The relationship between physical activity and health-related quality of life in adolescents: The MoMo study. *Journal of Physical Activity and Health*, 10(5), 708–715. <https://doi.org/10.1123/jpah.10.5.708>

Spinath, B., Stiensmeier-Pelster, J., Schöne, C., & Dickhäuser, O. (2002). *Skalen zur Erfassung der Lern- und Leistungsmotivation: SELLMO*. Hogrefe.

Spinath, F. M., & Wolf, H. (2006). CoSMoS and TwinPaW: Initial report on two new German twin studies. *Twin Research and Human Genetics*, 9(6), 787–790. <https://doi.org/10.1375/twin.9.6.787>

Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S., & Haag, N. (Eds.). (2017). *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich*. Waxmann.

Steinmayr, R., Crede, J., McElvany, N., & Wirthwein, L. (2016). Subjective well-being, test anxiety, academic achievement: Testing for reciprocal effects. *Frontiers in psychology*, 6, 1994. <https://doi.org/10.3389/fpsyg.2015.01994>

Stubbe, J. H., de Moor, M. H., Boomsma, D. I., & de Geus, E. J. (2007). The association between exercise participation and well-being: A co-twin study. *Preventive Medicine*, 44(2), 148–152. <https://doi.org/10.1016/j.ypmed.2006.09.002>

Stubbs, A. (2014). Reducing mental illness stigma in health care students and professionals: A review of the literature. *Australasian Psychiatry*, 22(6), 579–584. <https://doi.org/10.1177/1039856214556324>

Tikhodeyev, O. N., & Shcherbakova, O. V. (2019). The problem of non-shared environment in behavioral genetics. *Behavior Genetics*, 49(3), 259–269. <https://doi.org/10.1007/s10519-019-09950-1>

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews*

- Uhlig, J., Solga, H., & Schupp, J. (2009). Bildungsungleichheiten und blockierte Lernpotenziale: Welche Bedeutung hat die Persönlichkeitsstruktur für diesen Zusammenhang? [Inequality in educational opportunities: Underachievement and the role of personality traits]. *Zeitschrift für Soziologie*, 38(5), 418–440. <https://doi.org/10.1515/zfsoz-2009-0504>
- Valery, K. M., & Prouteau, A. (2020). Schizophrenia stigma in mental health professionals and associated factors: A systematic review. *Psychiatry Research*, 290, Article 113068. <https://doi.org/10.1016/j.psychres.2020.113068>
- van den Berg, S. M., Glas, C. A., & Boomsma, D. I. (2007). Variance decomposition using an IRT measurement model. *Behavior Genetics*, 37, 604–616. <https://doi.org/10.1007/s10519-007-9156-1>
- Varca, P. E., Shaffer, G. S., & Saunders, V. (1984). A longitudinal investigation of sport participation and life satisfaction. *Journal of Sport and Exercise Psychology*, 6(4), 440–447. <https://doi.org/10.1123/jsp.6.4.440>
- Veas Iniesta, A., López-López, J. A., Gilar Corbí, R., Miñano Pérez, P., & Castejón Costa, J. L. (2017). Differences in cognitive, motivational and contextual variables between under-achieving, normally-achieving, and over-achieving students: A mixed-effects analysis. *Psicothema*, 29(4), 533–538. <https://doi.org/10.7334/psicothema2016.283>
- Vilhjalmsson, R., & Thorlindsson, T. (1992). The integrative and physiological effects of sport participation: A study of adolescents. *The Sociological Quarterly*, 33(4), 637–647. <https://doi.org/10.1111/j.1533-8525.1992.tb00148.x>
- Vistorte, A. O. R., Ribeiro, W. S., Jaen, D., Jorge, M. R., Evans-Lacko, S., & Mari, J. D. J. (2018). Stigmatizing attitudes of primary care professionals towards people with mental disorders: A systematic review. *The International Journal of Psychiatry in Medicine*, 53(4), 317–338. <https://doi.org/10.1177/0091217418778620>
- Vitaro, F., Brendgen, M., & Arseneault, L. (2009). The discordant MZ-twin method: One step closer to the holy grail of causality. *International Journal of Behavioral Development*, 33(4), 376–382. <https://doi.org/10.1177/0165025409340805>
- Vlahovic-Stetic, V., Vidovic, V. V., & Arambasic, L. (1999). Motivational characteristics in mathematical achievement: A study of gifted high-achieving, gifted underachieving and non-gifted pupils. *High Ability Studies*, 10(1), 37–49. <https://doi.org/10.1080/1359813990100104>

- Volkow, N. D., Gordon, J. A., & Koob, G. F. (2021). Choosing appropriate language to reduce the stigma around mental illness and substance use disorders. *Neuropsychopharmacology*, 46(13), 2230–2232. <https://doi.org/10.1038/s41386-021-01069-4>
- von Stumm, S., & Plomin, R. (2018). Monozygotic twin differences in school performance are stable and systematic. *Developmental Science*, 21(6), Article e12694. <https://doi.org/10.1111/desc.12694>
- Waller, K., Kujala, U., Kaprio, J., Koskenvuo, M., & Rantanen, T. (2010). Effect of physical activity on health in twins: A 30-yr longitudinal study. *Medicine & Science in Sports & Exercise*, 42(4), 658–664. <https://doi.org/10.1249/MSS.0b013e3181bdeea3>
- Watson, A. C., Ottati, V., & Corrigan, P. (2003). From whence comes mental illness stigma? *International Journal of Social Psychiatry*, 49(2), 142–157. <https://doi.org/10.1177/0020764003049002007>
- Weiß, R., & Osterland, J. (2012). *Grundintelligenztest Skala 1 (CFT 1)* (6th ed.). Hogrefe.
- Weiß, R. H., & Weiß, B. (2006). *CFT 20-R. Grundintelligenztest Skala 2* (2nd ed.). Hogrefe.
- White, S. L., Graham, L. J., & Blaas, S. (2018). Why do we know so little about the factors associated with gifted underachievement? A systematic literature review. *Educational Research Review*, 24, 55–66. <https://doi.org/10.1016/j.edurev.2018.03.001>
- WHOQoL SRPB Group. (2006). A cross-cultural study of spirituality, religion, and personal beliefs as components of quality of life. *Social science & medicine*, 62(6), 1486–1497. <https://doi.org/10.1016/j.socscimed.2005.08.001>
- Wicker, P., & Frick, B. (2015). The relationship between intensity and duration of physical activity and subjective well-being. *The European Journal of Public Health*, 25(5), 868–872. <https://doi.org/10.1093/eurpub/ckv131>
- Willoughby, E. A., Polderman, T. J., & Boutwell, B. B. (2023). Behavioural genetics methods. *Nature Reviews Methods Primers*, 3(1), Article 10. <https://doi.org/10.1038/s43586-022-00191-x>
- Wu, H., Guo, Y., Yang, Y., Zhao, L., & Guo, C. (2021). A meta-analysis of the longitudinal relationship between academic self-concept and academic achievement. *Educational Psychology Review*, 33(4), 1749–1778. <https://doi.org/10.1007/s10648-021-09600-1>
- Young, A. I. (2019). Solving the missing heritability problem. *PLOS Genetics*, 15(6), Article e1008222. <https://doi.org/10.1371/journal.pgen.1008222>

- Zamir, A., Tickle, A., & Sabin-Farrell, R. (2022). A systematic review of the evidence relating to disclosure of psychological distress by mental health professionals within the workplace. *Journal of Clinical Psychology*, 78(9), 1712–1738. <https://doi.org/10.1002/jclp.23339>
- Ziegler, A., Ziegler, A., & Stoeger, H. (2012). Shortcomings of the IQ-based construct of underachievement. *Roeper Review*, 34(2), 123-132. <https://doi.org/10.1080/02783193.2012.660726>

Danksagung

Zum Abschluss der vorliegenden Arbeit möchte ich zunächst Prof. Dr. Frank Spinath danken, der mir die Möglichkeit zur Promotion eröffnete und diese trotz eines Umfeldes befristeter Verträge auch mit viel Planungssicherheit begleitete. Ohne die wertschätzende Unterstützung auch bei Hindernissen wäre die vorliegende Arbeit nicht möglich gewesen.

Das TwinLife-Projekt war eine hervorragende Forschungsbasis, für die ich mich nicht nur bei Prof. Spinath sowie Dr. Christoph Klatzka und Lena Paulus, sondern auch allen anderen Projektbeteiligten bedanken möchten.

Ebenso wertvoll war die Unterstützung innerhalb der Arbeitseinheit, angefangen beim klaglosen wandelnden Nachschlagwerk Dominik Weber. Auch allen weiteren aktuellen und früheren Kolleginnen und Kollegen bin ich dankbar, insbesondere Dr. Marco Koch, Dr. Nicolas Becker sowie Dr. Julie Levacher.

Unter den zahlreichen Studentinnen und Studenten, mit denen ich im Laufe der letzten Jahre mit sehr viel Freude zusammenarbeiten durfte, ist Lara Kubillus als Koautorin von Studie III besonders zu nennen.

Der letzte und wichtigste Dank geht an die Familie Dings, darunter insbesondere Frieda, die zum Glück nicht *nur* wegen des tollen Namens in die Familie einheiraten wollte.

Anhang

Studie I

Dings, A., & Spinath, F. M. (2021). Motivational and personality variables distinguish academic underachievers from high achievers, low achievers, and overachievers. *Social Psychology of Education*, 24(6), 1461-1485. <https://doi.org/10.1007/s11218-021-09659-2>

Studie II

Dings, A., & Spinath, F. M. (2024). Sports club participation impacts life satisfaction in adolescence: A twin study. *Psychology of sport and exercise*, 73, 102639. <https://doi.org/10.1016/j.psychsport.2024.102639>

Studie III

Dings, A., & Kubillus, L. K. (2025). Beyond case vignettes: do diagnostic labels affect how symptoms of mental illness are perceived? *Journal of Mental Health Disorders*, 5(1), 1-14. <https://doi.org/10.33696/mentalhealth.5.030>

Anmerkung: Studie III wurde ursprünglich beim *Journal of Mental Health Disorders* am 26. 12. 2024 zur Veröffentlichung akzeptiert, jedoch wurde aufgrund der nachträglich festgestellten Notwendigkeit einer Korrektur ein korrigiertes Manuskript eingereicht. Dieses wurde am 6. 8. 2025 bei dem gleichen Journal zur Veröffentlichung akzeptiert und befindet sich *in press*, wobei die Zitation voraussichtlich identisch bleiben wird. In dieser Version ist das Manuskript im folgenden Anhang wiedergegeben. Studie III folgt dabei aufgrund der Richtlinien des betreffenden Journals nicht der APA-, sondern der Vancouver-Zitierweise.

**Motivational and personality variables distinguish academic underachievers
from high achievers, low achievers, and overachievers**

Alexander Dings¹, Frank M. Spinath¹

Received: 25 February 2021 / Accepted: 10 August 2021 / Published online: 4 September 2021

Corresponding author: Alexander Dings, alexander.dings@uni-saarland.de

¹ Department of Psychology, Saarland University, Campus A1 3, 66123 Saarbrücken, Germany

Abstract

Previous studies have associated several variables concerning motivation and other domains with underachievement, i. e. a student's academic achievement falling short of what their cognitive abilities, as the best predictor of academic performance, would indicate. The present study extends these findings using a more rigorous approach in defining underachievers and suitable control groups. Using discriminant analysis, underachievers identified in a German twin family study were compared not only to achievers with comparable IQ scores, but also with students of lower aptitude, but comparably low grades, as well as overachieving students. Results confirm previous findings that compared to successful students, underachievers report lower levels of motivation and parental support; beyond this comparison, underachievers also differed from other low achievers, mostly in terms of their personality. In total, 40% of the variance between the groups were explained. Additionally, the data shed doubt on the common assumption that underachievers are an unusually heterogenous group of students.

1 Introduction

1.1 General background

General intelligence has long been known as a major predictor of academic achievement (Sternberg, Gigorenko & Bundy, 2001), with a recent meta-analysis reporting a median correlation of 0.54 (Roth, Becker, Romeyke, Schäfer, Domnick & Spinath, 2015). Yet even this substantial predictive validity, which was corrected for sampling error, error of measurement and range restriction, implies that about three quarters of the variance in the outcome are still left unexplained. A substantial body of research has therefore attempted to uncover other predictors of educational attainment.

There are several approaches to this question. Schneider and Preckel (2017) reported effect sizes for more than 100 predictors in a synthesis of meta-analyses, integrating about 3,300 single findings. While the resulting „ranking“ can hardly be taken literally as the interrelatedness between predictors is not taken into account, it illustrates clearly that influences on students' grades range across a variety of domains, such as personal traits, school variables, parental behavior, peers, etc. Other studies have therefore attempted to investigate which of the variables in this heterogeneous mix really matter. A straightforward approach for identifying predictors that matter beyond intelligence is to simply control for candidate variables statistically. Different constructs have been found to show some incremental validity, including traits like conscientiousness (Rimfeld, Kovas, Dale & Plomin, 2016), self-efficacy (Stankov & Lee, 2014) and self-control (Duckworth, Quinn & Tsukayama, 2012).

Underachievement research asks a similar, but not identical question to address some of the 75% of unexplained variance left by IQ. It focuses on those individuals for whom IQ turns out to be off the mark in predicting achievement. With a less than perfect correlation between IQ and achievement, it is self-evident that there will necessarily be a certain proportion of individual cases in which the actual outcome will differ substantially from the predicted one. Some students will receive very good grades despite unassuming scores on general intelligence tests; some students will show above-average cognitive abilities when tested, but not do as well in school. These latter students have been termed "underachievers" in educational psychology, and in particular in giftedness research (for an overview, see Reis & McCoach, 2000).

By focusing on the subgroup of students in which the best overall predictor of academic success is notably off the mark, this approach attempts to further qualify the more general findings cited above since validity coefficients obtained in a larger population do not necessarily translate into every student's individual achievement history. The question that underachievement research asks could thus be phrased as: In the unusual case that it is not (low) IQ that explains a case of low

achievement, what can explain it? It is not a given that factors known to be incrementally valid in general are the same that are at work in these “extreme” cases.

Previous studies have yielded a diverse range of findings on underachievement (see below). Its reverse case of overachievement has also been studied, albeit to a much lesser extent (Sparfeldt, Buch & Rost, 2010) – it might simply be a less pressing issue; also note that if a good IQ test result is used as a benchmark which the academic achievement of underachieving students is said to fall short of, you might just as well ask the reverse question for apparent “overachievers”: why did their IQ test performance fall short of what they achieve in school? Considering that grades are an aggregated measure of performance throughout a school year, a comparatively low result on a single IQ test might often be a fluke, resulting from random factors on that single test occasion.⁵ However, there are two issues in the conceptual and statistical approaches widely used in these studies that, in our view, entail validity problems: Underachievers are rarely compared to other low achievers, making it unclear whether their characteristics are actually specific to achieving lower *than expected*; and the most widely used approach to operationalize underachievement leads to questionable selections. This study is aimed at testing and extending previous findings on underachievement while addressing these issues.

1.2 Operationalizing underachievement

In attempting to operationalizing underachievement, researchers have taken several different approaches with different implications (Lau & Chan, 2001). Numerous authors stress the fact that there is no consensus on how to conceptualize and measure underachievement (e. g., Plewis, 1991; Hoover-Schultz, 2005; Preckel, Holling & Vock, 2006). The definition of underachievement provided by Reis and McCoach (2000), according to which underachievers are “students who exhibit a severe discrepancy between expected achievement (as measured by standardized achievement test scores or cognitive or intellectual ability assessments) and actual achievement (as measured by class grades and teacher evaluations)” (p. 157), can be seen as common ground, but it leaves much room for different approaches in defining a “severe discrepancy”. We will therefore describe the approach that seems most rigorous and valid to us.

Much of the previous literature adopted either the *difference method* or the *regression method* (s. Lau & Chan, 2001) to compare standardized measures of cognitive abilities and achievement scores – either by subtracting them from one another or, in order to avoid regression to the mean (s. Cone & Wilson, 1981), by calculating the residuals in a regression of the achievement measure on the

⁵ As far as overt difficulties in administering the test are concerned, IQ results are excluded from analyses in TwinLife for those cases where interviewers reported any issues, thereby reducing the likelihood of IQ results not reflecting actual abilities.

aptitude measure. The resulting difference or residuals are then interpreted as reflecting underachievement, and a cut-off is defined to select the most substantial cases. As a more recent, refined implementation of this rationale, Rasch models have been used, e. g. by Phillipson (2008). A different method, the *absolute split approach*, identifies underachievers categorically by defining cut-offs for low and high IQ as well as low and high grades, with the combination of high IQ and low grades representing underachievement. This approach might seem unfavorable at first glance because it dichotomizes the data and thereby discards some of the variance, but there are several advantages to it.

First of all, discarding some of the variance is not necessarily unwelcome. As alluded to in the introduction, the premise of underachievement research is to find out which (possibly unique) processes can explain those cases in which IQ as the most important predictor and academic achievement as the criterion do not align – in other words, extreme cases. The large middle ground of students who technically under- and overachieve to a small degree is not only of little interest in practice and hardly interpretable; these cases are also very prone to emerging simply due to random error variance, as the residuals in the regression method result from combining two variables with less than perfect reliability, making them rather noisy in a statistical sense (Zieger et al., 2012).

Second, the absolute split approach reflects a clear-cut definition of underachievement. As Hanses and Rost (1998) point out, this is questionable when using the regression approach, as it subsumes very different cases of partially unclear validity under the label of underachievement: it will „diagnose“ underachievement in many cases where the actual grades are just fine – but the student would have been expected to achieve even better results. While it can certainly be argued that this outcome might be a disappointment for the gifted student, we follow Sparfeldt, Schilling and Rost (1996) in reasoning that it would be dubious to use one and the same label for these students as for those whose unexpectedly low grades are actually critically low, and that an upper bound on grades is therefore advisable. Similarly, a lower bound for IQ is also advisable because it would be equally debatable to include students in the category of underachievement who were not expected to do all that well in the first place (such a lower bound is in place in most studies as they focused on gifted students). The absolute split method thereby focuses on those cases where there can be little disagreement about their validity (barring the possibility that they might have resulted from error variance): instances of underachievement are cases of academic failure (or being on the brink of failure) by students who score clearly above average in cognitive abilities. Note that such dramatic cases will be rarer in populations where the ability criterion and the achievement outcome are highly correlated; the regression method with 1 SD as the critical difference will, implausibly, always select about 15% of students as underachievers, regardless of whether IQ and grades

correlate at 0.2 or at 0.7 (Plewis, 1991)⁶.

Using the absolute split method offers the additional advantage of yielding convenient comparison groups (see Table 1): While the combination of high IQ and low grades represents underachievement, there are also groups combining high IQ and high grades (high achievers), low IQ and low grades (low achievers) and low IQ and high grades (overachievers).

Table 1 *Framework of the absolute split approach for defining underachievement and corresponding comparison groups*

	High IQ	Average IQ	Low IQ
High grades	(Expected) high achievers	<i>unselected</i>	Overachievers
Average grades	<i>unselected</i>	<i>unselected</i>	<i>unselected</i>
Low grades	Underachievers	<i>unselected</i>	(Expected) low achievers

This allows us to contrast underachievers with similar students differing either in grades or in IQ. Previous research mostly undertook only the first of these comparisons (i. e., the left column in Table 1) by focusing on gifted students and then contrasting underachievers with high achievers within this group, defined by the regression approach. This, of course, ensures that characteristics found in underachievers are not confounded with IQ (White, Graham & Blaas, 2018). However, as McCall and Lau (2000) pointed out, this leaves a similar issue unresolved - „virtually all of the research (...) confounds these correlates with poor grades“ (p. 787). In consequence, it remains unclear whether differences between achieving and underachieving smart students are actually attributable to *underachievement* rather than to low achievement in general. Low academic self-concept, for example, might generally be prevalent among students who receive bad grades – they might, almost trivially, simply report this lack of success when asked about their academic self-concept.

Comparing underachievers simultaneously in both directions – contrasting them to other smart students as well as other unsuccessful students – therefore seems necessary to identify which variables typically associated with underachievement, such as low academic self-concept, are actually specific to it. Using the label “underachievement” in research and in practice implies that these students are in some way different from low achievers in general, but this assumption has

⁶ To illustrate, imagine a fictitious population where a person's cognitive abilities and their academic achievement are highly correlated because all factors that can hinder a student to make full use of their abilities have been eliminated. As long as the correlation is not exactly 1, the regression method will still identify 15% of students as underachievers and thus suggest that the issue persists; in fact, such a misunderstanding can sometimes be found in the literature (e. g., Figg, Rogers, McCormick and Low, 2012) when it is suggested that underachievement „rates“ have stayed at an alarmingly high level, when this is in fact simply inevitable under the most prevalent definition, as pointed out by McCall (1994).

hardly been tested empirically. To our knowledge, only two studies have addressed the confounding of underachievement research results by grades, even though McCoach and Siegle (2003) mentioned it as a suggestion for future research. McCall and Lau (2000) did so by comparing underachievers to non-underachievers while using grades as a covariate, but underachievement was defined by the regression approach, incurring the aforementioned issue of subsuming very different cases under one label. Hanses and Rost (1998) included several groups in their analysis and were thereby partly able to differentiate the results along grades and IQ, but their comparison groups included average achievers rather than low achievers. Other studies included non-gifted students as a control group, but did not differentiate them by grades (e. g., Vlahovic-Stetic, Vidosic and Arambasic, 1999; Dixon, Craven & Martin, 2006).

1.3 Empirical background

The research detailing the differences between underachievers and other students – usually gifted achievers – is diverse and sometimes contradictory (Reis & McCoach, 2000; Peterson & Colangelo, 1996; White et al., 2018) to such an extent that Figg, Rogers, McCornick, and Low (2012) conclude that there is „very little, if any, consensus on any answers“ (p. 54).

Some longitudinal research gives credence to the belief that the basic premise of underachievement research is valid: it is not mainly error variance that studies are trying to explain; underachievement is a meaningful phenomenon with a possibly lasting impact. In other words, temporary underachievement does not simply “regress to the mean” in most cases, as one would expect if the discrepancies between grades and cognitive abilities were entirely the result of measurement error. For example, Colangelo, Kerr, Christensen and Maxey (1993) reported lower college admission test scores for high-school underachievers (although to a lesser extent than their underachievement would have suggested, indicating that some regression the mean did occur). Studies by Johnson, McGue and Iacono (2006) and Peterson (2000) also indicate that at least some cases of underachievement persist longitudinally. On the other hand, all of these results also demonstrate that underachievement is clearly not an irreversible “condition”, as many students get back on track later on (see also Uno, Mortimer, Kim & Vuolo, 2010). Support from family and peers, as well as persistence in setting ambitious goals despite the difficulties, appear to be crucial factors for this to happen (Peterson, 2000; Hwang, Lee, Lim, Seon, Hutchison & Pope, 2014; McCall, 1994), making it all the more important to recognize underachievement as such and not mistake it for an unalterable lack of ability. This is not a trivial task: Rost and Hanses (1997) found that teachers in their study could not identify underachieving students in their classes very well; in a more recent study, teachers had only a moderately accurate impression of their students' attitudes towards school

- in particular, they often underestimated their students' goal valuation (Siegle, DaVia Rubenstein and McCoach, 2020).

There is no clear overall picture on what determines underachievement, partly due to a relative paucity of comparable studies (White et al., 2018). Different procedural explanations for the development of underachievement have been suggested (e.g., Snyder & Linnenbrink-Garcia, 2013), but the empirical basis for them is scarce. The same is true for various “typologies” of underachievement suggested by practitioners (see Reis & McCoach, 2000). Nothing demonstrates this need for clarification better than the fact that interventions aiming to reduce underachievement have so far met with limited success overall, according to a recent meta-analysis (Steenbergen-Hu, Olszewski-Kubilius & Calvert, 2020).

In the current study, we focus on two particular domains: motivation and personality.

1.3.1 Motivation

Motivational variables are at the core of one of the few theoretical models that attempt to provide a general explanation for at least some cases of underachievement. The Achievement Orientation Model (or AOM; Siegle, McCoach & Roberts, 2017), which is based on existing theories of motivation, hypothesizes that students' motivational attitudes toward achievement serve as a necessary link between their abilities and actual success. In particular, (gifted) students need to believe that trying to be successful is a worthwhile goal (goal valuation); they need to believe in their ability to do so (self-efficacy); and they need to feel supported in their efforts by their environment, such as parents or teachers (environmental perceptions). The AOM presumes that these components cannot compensate for each other: each of them needs to be fulfilled at least to a minimal degree for the student to be successful. If this motivational basis is present, a behavioral component summarized as self-regulation is further necessary to successfully put the motivation to succeed into practice; if the student does not know how to learn effectively, for example, their abilities and motivation will not translate into success.

It has repeatedly been found that the constructs referred to in the AOM are implicated in underachievement in the sense that compared (only) to high achievers, underachievers report lower academic self-concept, goal orientation and support from their parents (Vlahovic-Stetic et al., 1999; McCoach & Siegle, 2003; Balduf, 2009; Landis & Reschly, 2013; White et al., 2018; Mofield & Parker Peters, 2019). The AOM's other notion – that motivation must be complemented by skills and strategies that enable make it possible to learn efficiently – is also supported by empirical evidence. Hyperactivity and other kinds of dysfunctional behavior is an antecedent to

underachievement in kids (as well as a consequence of it; Hinshaw, 1992). Castejon, Gilar, Veas & Miñano (2016) found lower use of several kinds of learning strategies among underachieving students. Students may be well aware of this, as a majority of underachieving gifted students in Colangelo et al.'s (1993) study stated they needed to improve their learning strategies (as opposed to just 15% of high achieving students; furthermore, this difference was somewhat specific, as no such gap appeared on questions asking whether students needed help with educational plans or with personal problems). The finding that learning strategies mediate the association between parental support and achievement is also in line with the AOM (Veas, Castejón, O'Reilly & Ziegler, 2018). Direct tests of the model have also provided support for it (Ritchotte, Matthews & Flowers, 2014; Rubenstein, Siegle, Reis, McCoach & Burton, 2012). In-depth interviews with a small sample of gifted students confirmed that the constructs in the AOM are subjectively important for these students as well (Barbier, Donsche & Verschueren, 2019). The motivational variables used in the current study do not match these components so exactly as to constitute a direct test of the AOM, but they do represent the variety of motivational components summarized in the model. Furthermore, we tested whether it is true that these components cannot compensate for one another, as this assumption of the AOM currently lacks empirical validation⁷.

1.3.2 Personality

We chose the Big 5 personality variables (Costa & McCrae, 1992) as another branch of predictors because there is surprisingly little empirical research on how underachievers differ (or do not differ) from other students in this major domain of interindividual differences. In general, the Big Five do not show uniform effects on academic achievement, with the exception of a small positive association with conscientiousness (O'Connor & Paunonen, 2007); since most underachievement research compares underachievers with high achievers only and thus basically looks for correlations with achievement (within the group of high-IQ students), this overall null correlation might have led researchers to regard the Big Five as an unattractive object of study. However, the question remains whether the general null results can be extrapolated to the particular subgroup of underachievers, especially under the framework of the present study. The findings that do exist, as listed in Reis and McCoach (2000)'s review, are diverse and contradictory; almost no studies have since helped to clarify the situation, with the exception of Preckel and colleagues (2006) who reported that underachievers displayed lower conscientiousness; they did not report testing the other Big Five dimensions.

⁷ Ritchotte et al. (2014) reported support for it, but the test they reported just reiterated the main effect of the motivational components; it did not test for compensation, which concerns the interaction between the components.

It is often stressed that underachievers are a very heterogeneous group (Siegle, 2018). This might be one explanation for the diversity of findings on the topic, but it might also partly be due to the different empirical approaches. However, to our knowledge, McCoach and Siegle's (2003) finding of larger variances in a group of underachievers than in a group of high achievers (concerning motivational attitudes) as well as a rather qualitative finding by Smith (2007) constitute the only reported evidence of the conjured larger heterogeneity of underachievers in contrast to other groups. We therefore tested whether such a pattern held in our sample as well.

2 Materials and methods

The present study drew on a large sample collected in TwinLife, a panel study that surveys pairs of twins as well as their extended families in Germany, with a focus on following and explaining different life courses (for an overview, see Mönkediek et al., 2016). The sample is mostly representative for the German population, although families with higher socio-economic status are somewhat overrepresented (Lang & Kottwitz, 2017). Data from the first two data collection phases are available as scientific use files via the GESIS data archive (<https://doi.org/10.4232/1.13208>). As one aspect of tracking a person's road to success in life, TwinLife includes a number of motivational and environmental variables pertaining to academic success. Some of these self-report scales were used as approximations of the AOM constructs in this study: learning valuation (meant to resemble goal valuation in the AOM); self-efficacy and academic self-concept (self-efficacy in the AOM); and parenting style (reported by the children), which consisted of scales measuring parental support, autonomy and control (environmental perceptions). TwinLife did not include a direct measure of self-regulation in the first data wave (which was used here); our proxy for successful (or unsuccessful, rather) self-regulatory behavior necessary as a further link to academic achievement according to the AOM therefore was hyperactivity (a subscale of measuring externalizing behavior).

After four joint years of elementary school, Germany's school system is stratified, with grammar schools (Gymnasium) being the highest level in the sense that the degree obtainable in these schools (Abitur) is a prerequisite for studying at a German university and the most valuable for the labour market. This creates an issue when trying to study underachievement: A student can be successful or unsuccessful in terms of better or worse grades within the school that they attend, but educational attainment in the long run is also characterized by what level of school the student attends in the first place (see Uhlig, Solga & Schupp, 2009, whose study looks at underachievement through this alternative perspective). As the latter variable (the different school levels in the stratified system) constitutes merely an ordinal scale, a reliable metric to compare good grades in a lower-level school

to bad grades in grammar school is not feasible. The analysis was therefore restricted to over- and underachievement within grammar schools. Students' grades were collected by photographing the most recent report card (these are issued twice a year). If taking a photo was not possible or not agreed to, parents were asked to provide the grades for German and Math (this was the case for only about 1% of the sample). When analyzing German school grades, it is common practice to only gather data for the subjects of German and Math, as these are the only two subjects that are taught in virtually every school. This procedure was followed in the present study as well; when report cards were not available, parents were asked to report (only) on these two subjects. All report cards dated from the same year as the general survey used for the other variables, and students were excluded from the sample if they were no longer attending school at the time of the survey (e.g., their most recent report card was their final one). With these restrictions, a total of 612 students (324 female; mean age = 14.6 years, SD = 2.8) remained in the sample. They came from 377 different families and included 103 complete monozygotic (MZ) pairs, 104 dizygotic (DZ) pairs, 36 twins from incomplete MZ pairs (because their co-twin did not attend a gymnasium school, for example); 75 twins from incomplete DZ pairs, and 87 siblings of twins.

General intelligence was measured using a short version of the Culture Fair Test, or CFT (Weiß & Osterland, 2012), administered via computer and consisting of four subtests measuring reasoning, figural reasoning, figural classification and matrices. It is a frequently used (s. Kuhn, Holling & Freund, 2008) test aiming to measure fluid intelligence (i. e., capability of solving problems and processing complex information, Cattell, 1963) through tasks that are designed to be as “culturally fair” as possible in that they rely on basic, universal contents such as geometric patterns, as opposed to specific knowledge that depends more strongly on an individual's culture and language. A factor analysis of the four scales yielded a clear unifactorial solution. Factor scores were used as an overall index of cognitive ability (although this did not really make a difference as they were very highly correlated with simple sum scores). Great care was taken to exclude (possibly) invalid cases by having home interviewers note any irregularities as well as by having two independent raters judge all cases with suspicious data patterns, i. e. subtests diverging strongly from one another; for further details, see Gottschling (2017).

Both variables (grades as well as cognitive test scores) were standardized by age and sex (as is recommended when using twin data; McGue & Bouchard, 1984).⁸

The instruments used in TwinLife to measure the predictor variables in this study are as follows.

⁸ It was also checked whether grade averages differed substantially across Germany's 16 federal states; while statistically significant, the differences were small enough in effect size, with $\eta^2 = 0.053$, that standardizing grades by state was deemed likely to be spurious rather than helpful: some small states were only represented by a handful of students and it is unknown whether the differences are even attributable to unwelcome method variance.

Self-efficacy, learning motivation and academic self-concept were measured using existing German-language short scales (Beierlein, Kovaleva, Kemper & Rammstedt, 2012 for self-efficacy; Spinath, Stiensmeier-Pelster, Schöne & Dickhäuser, 2002, for learning motivation; Dickhäuser, Schöne, Spinath & Stiensmeier-Pelster, 2002, for academic self-concept). Cronbach's alpha coefficients in our sample were 0.71, 0.70 and 0.72, respectively. Sample items include “I can rely on my own abilities in difficult situations” for self-efficacy and “In school, I want to learn as much as possible” for learning motivation. Participants indicated how well these statements described them on a five-point scale ranging from “doesn't apply at all” to “applies exactly”. For academic self concept, children were asked to complete an item such as “I am ... for school” with options ranging from, for example, “not talented” to “very talented”.

Parenting style was adapted from Spinath and Wolf (2006) and measured parental support, parental autonomy and parental control with three items each; these subscales reached reliabilities between 0.68 and 0.79. Participants indicated the extent of their agreement to statements such as “My parents console me and help me when I have problems in school” (support), “When my parents help me with my studies they encourage me to find the solution myself” (autonomy) and “When I get a poor grade, my parents complain and demand that I work harder” (control).

Hyperactivity was part of a scale on externalizing behavior adapted from Goodman, Meltzer and Bailey (1998). Cronbach's alpha for the five items was 0.72. An example item is “I am easily distracted, I find it difficult to concentrate”. The Big Five personality subscales of extraversion, neuroticism, agreeableness, openness to experience and conscientiousness were assessed using the Big-Five Inventory SOEP (s. Gerlitz & Schupp, 2005). Participants were asked to indicate how strongly (on a 7-point scale in this case) they agreed to statements about themselves, all of which began “I see myself as someone who...” and then mentioned various traits typical for each of the five factors, e. g., “...worries a lot” for neuroticism or “...tends to be lazy” for (low) conscientiousness. These mini scales with 3-4 items each achieved Cronbach's alphas ranging from 0.55 to 0.74.

Data were collected through home visits by trained interviewers. Because of the somewhat sensitive topics, most of these measures were assessed through self-administered computerized testing; for learning motivation and academic self-concept, however, the interviewers asked the questions directly and recorded the answers. More detailed information, including all of the items, is available from the TwinLife Scales Manual (Baum, Klatzka, Iser & Hahn, 2020).

As explained in the introductory section, students were classified as expected low and high achievers or as unexpected low and high achievers (i.e., under- and overachievers) by selecting those students who combined a notably above- or below-average IQ with above- or below-average

grades. Of course, specifying the cutoff for "notably" above or below-average scores is an arbitrary choice; there is a tradeoff between the sample size that is retained after the selection and the validity and significance of the cases. Too lax a criterion will result in more spurious cases due to error variance, as well as cases that can technically be termed, for example, underachievement, but do not actually represent severe difficulties in school (see section 1.2). Cutting off both the standardized intelligence scores and the standardized grade averages derived earlier at Z values of 0.5 above and below the mean resulted in a reasonable balance between sample size and validity concerns (see below). Additionally, although defining cutoffs is in principle an arbitrary decision, school grades do offer some reference points. Grades in the German school system range from 1 to 6, with 1 being the best grade. An important cutoff lies between 4 and 5, as a grade of 4 or better "is generally required in each of the subjects that have a bearing on promotion" (Lohmar & Eckhardt, 2014). Therefore, there can be little doubt that a grade of 5 or 6 would definitely fall below a reasonable cutoff for low achievement. However, these grades are so rare (about 2% of the cases in our sample) that selecting only students with a 5 or 6 in either German or Math would have reduced the sample size drastically. In fact, the distribution of grades is skewed towards good grades so strongly that a few students were classified as underachievers or low achievers with a cutoff of 0.5 SD while receiving a 3 ("satisfactory") in both subjects, despite the fact that this selection was already much more restrictive than the usual regression approach. These cases were excluded in order to ensure that all low achievers and underachievers were at least on the brink of failing a class. Despite our already relatively strict selection approach, there were still 9 such cases out of originally 30 underachievers and 19 out of 82 low achievers. As far as our German sample can be generalized, and as far as readers follow our interpretation of „serious“ underachievement, it thus appears to be a remarkably rare phenomenon: In total, there were 352 different students who fell either into the high IQ group or the low achievement group, or both; the latter case, the overlap between both categories, represents underachievement, but happened only in 21 cases out of the 352.

In total, 21 underachievers, 30 overachievers, 63 low achievers and 113 high achievers remained after these selection procedures (the uneven distribution of low and high achievers is partly due to the exclusion of some low achievers as mentioned above and partly due to the categorical nature of the achievement variable). As the sample from which these students were selected included 612 students, 63% were not classified into any of the extreme groups as either their IQ or their grades, or both, were more or less average. A total of 3% and 5% were identified as underachievers or overachievers, respectively. Note that the usual approach of using a difference of 1 SD between standardized grades and IQ scores results in about 15% selected cases in either direction, but, in our view, will include many cases whose classification as practically meaningful underachievement is

doubtful. That our approach was comparatively strict can also be seen from the fact that underachievers represented about 10% of high-IQ students in our sample; in a literature review by White et al. (2018), this figure ranged from 9% to 28% in different studies (with different methodologies and samples).

The 227 students came from 182 different families and included 17 complete MZ pairs and 20 complete DZ pairs. No family included two underachievers, while three families included two overachievers each (MZ twin pairs in two cases, siblings in one case). The average age was 14.6 years ($SD = 2.8$); due to the nature of the panel study, which consisted of cohorts separated by about 6 years, a majority of students were either 11 years old (99 cases) or 17 years old (72 cases). Among the predictor variables used in the discriminant analysis, 2.4% of the values were missing. Little's MCAR test, a conservative test assessing whether data are missing completely at random (Little, 1988), was not significant ($\chi^2 = 124.13$, $df = 104$, $p = 0.09$), indicating that it was sufficient to replace missing data using the Expectation Maximization procedure rather than using the more complex approach of multiple imputation (Tabachnick & Fidell, 2011).

3 Results

Table 2 displays zero-order correlations between the predictor variables within the final sample, i. e. those students that were classified as either underachievers, overachievers, or belonging to a suitable control group; descriptive statistics are shown in Table 3, grouped by achievement category. Results of the discriminant analysis are shown in Tables 4 and 5.

Table 2 Zero-order correlations for the predictor variables in the different achievement groups

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
(1) Parental support											
(2) Parental autonomy	.79**										
(3) Parental control	-.30**	-.22*									
(4) Learning motivation	.22*	.25*	-.17*								
(5) Self-efficacy	.16*	.23*	-.17*	.40**							
(6) Academic self-concept	.21*	.17*	.08	.21*	.26**						
(7) Hyperactivity	-.10	-.13	.32**	-.23*	-.27**	-.27**					
(8) Neuroticism	-.09	-.14	-.10	.11	-.21*	.03	.01				
(9) Extraversion	.26**	-.16*	-.06	.06	.23*	.18*	.03	-.21*			
(10) Conscientiousness	.21*	.22*	-.31**	.33**	.34**	.29**	-.44**	.12	-.02		
(11) Agreeableness	.28**	.25**	.19*	.24**	.04	.21*	-.23*	.20*	.06	.29**	
(12) Openness	.14*	.06	-.07	.27**	.13	.09	-.12	-.02	.24**	.12	.23*

* $p < 0.05$. ** $p < 0.01$.

Table 3 Group means and standard deviations for the predictor variables in the different achievement groups

	Underachievers (n = 21)	Low achievers (n = 63)	High achievers (n = 113)	Overachievers (n = 30)
Parental support	9.24 (2.12)	8.86 (2.74)	9.75 (2.27)	9.61 (2.71)
Parental autonomy	8.71 (2.63)	9.03 (2.61)	9.53 (2.41)	10.05 (2.37)
Parental control	4.71 (2.12)	4.79 (3.22)	2.99 (3.03)	2.73 (2.75)
Learning motivation	8.81 (1.83)	8.84 (1.88)	9.54 (1.61)	10.14 (1.63)
Self-efficacy	7.86 (1.93)	8.37 (2.08)	9.18 (1.50)	9.47 (1.76)
Academic self-concept	6.58 (0.67)	6.83 (1.33)	7.52 (1.06)	7.57 (0.77)
Hyperactivity	3.71 (2.05)	3.88 (2.41)	2.58 (1.93)	3.05 (2.18)
Neuroticism	11.86 (4.33)	11.71 (5.11)	11.17 (4.68)	11.17 (4.01)
Extraversion	10.57 (5.04)	12.22 (4.20)	11.27 (3.67)	12.27 (3.62)
Conscientiousness	10.00 (2.92)	11.16 (2.96)	12.74 (2.82)	13.43 (3.17)
Agreeableness	13.00 (2.05)	13.56 (3.44)	13.69 (2.89)	14.57 (2.53)
Openness	17.19 (4.26)	16.06 (4.06)	17.65 (3.97)	16.43 (3.97)

Note. Variable ranges were: 0-12 for the parental involvement variables, learning motivation, self-efficacy, and academic self-concept; 0-10 for hyperactivity; 0-18 for E, C and A, and 0-24 for N and O.

Table 4 Values of the functions at group centroids for the two first discriminant functions

	Function 1	Function 2
Underachievers	-.79	-.73
Low achievers	-.75	.25
High achievers	.45	-.21
Overachievers	.44	.77

Note. The third discriminant function is omitted because of its lack of significance.

Table 5 Canonical correlations of each predictor with the first two discriminant functions as well as (in brackets) weights in the discriminant functions

	Function 1	Function 2
Parental support	.25 (.06)	-.10 (-.21)
Parental autonomy	.23 (-.04)	.18 (.10)
Parental control	-.52 (-.48)	-.04 (.30)
Learning motivation	.40 (.15)	.22 (-.87)
Self-efficacy	.48 (.22)	.22 (.29)
Academic self-concept	.58 (.57)	.12 (.19)
Hyperactivity	-.47 (-.03)	-.22 (-.25)
Neuroticism	-.10 (-.26)	-.01 (.20)
Extraversion	-.07 (-.38)	.33 (.34)
Conscientiousness	.58 (.28)	.31 (.37)
Agreeableness	.13 (-.17)	.26 (-.20)
Openness	.22 (.21)	-.30 (.14)

Box's M test was insignificant (Box's M = 286.8, $p = 0.357$), indicating that the assumption of homogeneity of covariances necessary for discriminant analysis was not violated. The first discriminant function explained 66% of the variance (canonical $R^2 = 0.26$), the second function explained 32% (canonical $R^2 = 0.14$); the contribution of the third function (3% of the variance, canonical $R^2 = 0.01$) was insignificant. Wilks's lambda for all functions combined was $\Lambda = .63$, $X^2(36) = 101.4$, $p < .001$; over and above the first function, the other two still differentiated the groups significantly, with $\Lambda = .85$, $X^2(22) = 36.8$, $p = 0.025$.

The position of the group centroids on the first function suggest that it almost exclusively separated successful students from unsuccessful ones: Underachievers (group centroid at -0.79) were grouped together with low achievers (-0.75), while overachievers (0.44) and high achievers (0.45) were indistinguishable at the other end. In other words, as far as underachievers are concerned, this function highlighted those variables that differentiated underachievers from successful students, regardless of whether those are high achievers or overachievers.

The canonical correlations indicate that the successful students reported a higher academic self-concept, higher self-efficacy, learning motivation and conscientiousness; they reported less hyperactivity and a more functional style of parental involvement (more autonomy and support, less control). Note that discriminant analysis yields two different outputs; apart from the canonical correlations, the weights in the discriminant functions themselves indicate each variable's unique contribution to the empirical separation of the groups. In this regard, not all of the variables in which the two groups differed were crucial for distinguishing them (for example, if all the other variables are known, hyperactivity has little incremental value beyond them). Academic self-concept was the most prominent separating variable.

The group centroids of the second function indicate that, after separating successful and unsuccessful students, classification can mostly be improved by contrasting low-IQ and high-students, and in particular by contrasting underachieving students (group centroid at -.73) to overachieving students (0.77). Expected low (.25) and high achievers (-.21) were separated to a lesser degree, but were more similar to the students who were in the same IQ group as they were.

In other words, this function differentiates whether a student over- or underperforms in terms of IQ, and it indicates which aspects are associated with underachievement (and its mirror image of overachievement) rather than just with low achievement in general (as that part of the variance has already been dealt with in the first function).

The explanatory value of this function was, while significant, rather small (explaining 11% of the overall variance). The most substantial canonical correlations were present among the Big Five variables, with underachievers reporting lower (and overachievers higher) extraversion and conscientiousness, as well as higher openness to experience. Motivational variables generally pointed into the direction of a more dysfunctional pattern in underachievers (lower motivation and belief in themselves, higher hyperactivity, less parental support), albeit with only small coefficients.

The group means (see Table 1) offer tentative support for interpreting the results as a meaningful pattern rather than random variance, as the group differences emerged about equally in both directions in all of the notably contributing variables – overachievers mirrored underachievers in their differences from the comparison groups.

Using the discriminant functions, 45.8% percent of the cases could be classified correctly (as opposed to 25% by chance alone). However, this value dropped to 36.1% when each case was excluded from predicting itself (leave-one-out cross-classification; Molinaro, Simon & Pfeiffer, 2005) pointing to some likely overfitting in the results. Using a pre-defined group size is perhaps a more realistic approach, as the expected proportion of underachievers under a given selection criteria is known by definition; on the other hand, this means that mostly high achievers and low achievers are predicted, not the more „interesting“ cases of under- or overachievement. This approach yielded a correct classification of 59.0% without and 53.3% with cross-classification.

Based on the variables used for classifications, underachievers appeared to resemble (non-underachieving) low achievers (their likelihood of belonging to that group was estimated at 26% on average) somewhat more closely than high achievers (20%). This analysis was repeated with academic self-concept excluded because it could be argued that underachievers resemble other low achievers simply because they have basically stated (indirectly) that they are low achievers (although it should be noted that academic self-concept and grades correlated only moderately at $r = 0.36$). The difference shrunk to 25% vs. 23%. A similar yet reverse pattern was apparent for overachievers, who were more similar to expected high achievers (with comparable grades) than to expected low achievers (with comparable IQ) with values of 26% vs. 20% when including and 25% vs. 21% when excluding academic self-concept. Another way of operationalizing the similarity of the groups is to calculate the distance between the group centroids; at 0.97, the distance between underachievers and low achievers was smaller than between underachievers and high achievers (1.34).

In order to test the hypothesis that underachievers are more heterogeneous as a group than the other

“types” of students, a Levene test was performed on all variables as well as the discriminant scores for both functions. Variances for the overall discriminant scores did not differ significantly; among the individual variables, self-efficacy, academic self-concept, agreeableness and extraversion displayed heterogeneous variances across groups (not accounting for multiple testing). However, descriptive values (see Table 2) show that instead of underachievers, it is the group of other “expected” low achievers that displayed the highest variance in these variables, except for extraversion. In academic self-concept and agreeableness, underachievers actually displayed the lowest variance among all four groups. Therefore, we found no evidence to support the assumption that there is more variation among underachievers in the motivational and personality variables examined here – if anything, the contrary is true in our sample.

The assumption made in the AOM that the motivational precursors to achievement cannot be compensated for among each other was tested by checking whether the minimum among these (standardized) variables, as well as the number of variables below a certain low threshold (1 SD below the mean, approximately following Matthews et al., 2014), has any incremental validity in predicting underachievement when the mean level of these variables is taken into account. Underachievement was operationalized via the regression approach (see introduction) in this instance⁹. As expected, the mean Z-scores in the variables that most closely resembled the AOM components (learning motivation, self-efficacy, academic self-concept and parental support as the closest proxy to environmental support) correlated substantially with the underachievement score ($r = 0.37$, $p < 0.001$; the coefficient is positive because higher scores in the regression approach denote overachievement). Their minimum value as well as the number of components below the threshold correlated similarly highly, and in plausible directions, with the underachievement score ($r = 0.27$ and $r = -0.37$, respectively; p 's < 0.001). However, when the mean was introduced as a control variable in a partial correlation, both associations became insignificant ($r = -0.09$, $p = 0.17$ for the minimum value and $r = -0.11$, $p = 0.08$ for the number of low scores). Thus, the non-compensation hypothesis was not supported.

4 Discussion

4.1 General Discussion

The present study approached the issue of underachieving students under a slightly extended framework than previous research. Motivational variables similar to those in the Achievement

⁹ This was done in order to test as liberally as possible (by not discarding any variance in the data) so that a possible null result might not be explainable by too conservative a test.

Orientation Model as well as the Big Five personality variables were examined as traits that might differentiate underachievers from other students. This was in part a replication of established differences between underachievers and high achievers; adding the simultaneous comparison to low achievers (and overachievers) as well as, in terms of variables used, the Big Five, were the novel parts of the study. The first discriminant function, separating both kinds of successful students from both kinds of unsuccessful ones, indeed reiterates the point that motivational variables are strongly associated with academic achievement. Correlations between achievement and a functional parenting style were also replicated. The Achievement Orientation Model as a framework for understanding what can get in the way of a gifted student translating their cognitive abilities into educational attainment thus receives some support from our findings.

The second point is somewhat less clear; while the second function in the discriminant analysis significantly improved classification, coefficients were generally small (and possibly unstable, see below).

However, with these restrictions in mind, our study adds to previous research in justifying the construct itself: It appears that underachievers do in fact differ not only from other highly intelligent students, as previous research had already demonstrated, but that some differences to other low achievers are also discernible, albeit smaller. Yet this difference is only half the story: the discriminant analysis included four groups, and overachievers appeared to be more or less a mirror image of underachievers. This corroborates a so far very limited research body on overachievement (Sparfeldt et al., 2010). The bottom line is that those relatively rare cases where cognitive abilities – as the most important predictor of academic achievement – are “overruled” by other factors are not solely the product of random (error) variance; those cases instead appear to have something in common.

According to our results, these things in common are, among others, personality differences. It is a somewhat novel finding that the specific discrimination between underachievers and low achievers (as well as overachievers and high achievers; a parallel we will not mention explicitly hereafter for the sake of brevity) was in part due to the Big Five factors. Low agreeableness, low conscientiousness and low extraversion were typical for our sample of strongly underachieving students; further research might corroborate or disprove these tentative findings. The finding that the underachievers in our sample tended to be introverted could help explain the findings cited above (Rost & Hanses, 1997, and Siegle et al., 2020) which indicate that teachers often have trouble recognizing which of their students underachieve, and why.

The motivational variables exerted less influence in discriminating underachievers and low achievers. Although the variables mostly pointed in the expected direction numerically, it cannot be

denied that academic self-concept in particular was not very useful for identifying underachievers beyond the function that separated the successful from the unsuccessful groups. The finding that underachievers scored lower on academic self-concept than successful students is not surprising as this merely reflects the obvious facts on their report cards; the plausible conjecture that their self-concept might be even lower than that of other low achievers (because they might compare it to a higher standard, suspecting that they “should” be able to do better) could not be confirmed. (Note that, as the descriptive values in Table 2 show, this was not due to a floor effect.) Low scores on the more generally phrased items assessing self-efficacy were somewhat more typical for underachievers, but only slightly. When trying to identify students who have "hidden potential" that has not been brought to life yet, it might seem reasonable to think that the most disheartened students may be those who are falling short of their potential because they just are not believing in themselves. However, not believing in their capabilities seems to be almost as typical for any unsuccessful student as for underachievers. In this light, it is unsurprising that an intervention targeting this aspect was unsuccessful (Obergriesser & Stoeger, 2015) as it might have tackled a symptom rather than a cause. Note that this does not necessarily mean that underachievers do not know they are underachieving. Students were asked about their practical ability to succeed, not about their intelligence (or the relation between these two constructs).

Besides not being identical to either other low achievers or other highly intelligent students, detailed results of the discriminant analysis indicated that in terms of the motivational and personality variables used here, underachievers are hardly even closer to one group than the other – they cannot clearly be put into either box. They are somewhere in between, a result echoing findings by Hanses and Rost (1998) in one of the few other studies that included similar comparisons.

4.2 Limitations

Although our results do provide some indirect support the AOM, we found no evidence for its assumption that if one of the different facets of a student's motivation is particularly low, they are more likely to underachieve than a student with the same mean level of motivation, but a more balanced distribution. That being said, the variables we used only approximate those in the AOM. In the sample used in this study, there was also no indication for another suggestion that can frequently be found in the literature on underachievement - that they are more heterogenous than other groups of students. If anything, our results point to the contrary. Further empirical clarification might be warranted, especially because the way we tested it, using variance as a proxy for heterogeneity, certainly does not answer the question once and for all. It does not follow, of

course, that underachievement is a uniform phenomenon.

One obvious caveat of the study is that no causal inference can be made; for example, it remains unclear whether an unhelpful manner of parental involvement actually causes students' difficulties - parents might also simply be reacting to their children's bad grades.

The results also warrant some replicability concerns. In balancing the tradeoff between sample size and robustness of the selection, we focused on validity, which necessarily diminished the potentially large sample size to only 51 overachievers and underachievers in total. With ten predictors in the discriminant analysis, a sample size of 21 in the smallest group can only offer tentative results, even though the comparison groups were larger. The fact that classification worsened notably when leave-one-out classification was used suggests that some non-negligible overfitting was indeed taking place. Less strict inclusion criteria would have retained a larger part of the original sample and therefore allowed for more reliable conclusions, but their validity would have been doubtful. Part of the problem was the fact that bad grades (4-6 out of the 1-6 range in the German school system) were quite uncommon in the sample, perhaps reflecting a general tendency of grade “inflation” in Germany (Grözingen & Baillet, 2015). Consequently, students with clear-cut underachievement are actually relatively rare. An argument could be made that underachievement can also entail achieving mediocre grades despite great expectations – especially as the aforementioned inflation slowly changes the relative meaning of these grades. But even if we had followed this logic, including these cases in our analysis would have run the risk of grouping very different cases together under one label and thereby comparing apples and oranges. Besides the sample size, it must also be noted that due to the nature of the panel study, short scales were used to measure the predictor variables; more reliable instruments might yield more stable results.

The fact that we used a sample from a twin study also warrants some caution in interpreting the results as the underlying observations are not fully independent of each other (Carlin, Gurrin, Sterne, Morley & Dwyer, 2005). In order to maintain a reasonable sample size, we decided against including only one randomly selected member of each twin pair of family. The sample used for the final analysis ended up including 20 MZ and 17 DZ pairs; still, due to the selection process of filtering out students who did not fit into any of the achievement groups we were interested in, the majority (60%) were students who were the sole member of their family to be included. Note also that the 21 underachievers actually came from 21 different families - if their twin or sibling was part of the sample, they were in other groups. Among the 30 overachievers, there were two MZ pairs and one sibling pair. Overall, this does not give the impression that the differences between the groups could be explained by artificially increased similarity within them. On a sidenote, the fact that the underachievement and overachievement groups do *not* consist of notably many (MZ) pairs

could be taken as a tentative indication that the heritability of over- or underachieving, as opposed to academic achievement per se, is not strikingly high. Exploring this issue was not an aim of this paper, but could be a question for future research.

On the other hand, it is notable that the study yielded meaningful results even though its range was necessarily restricted because it was limited to grammar school students. When a wider perspective is taken, other factors might come into play as well. As shown by Uhlig et al. (2009), socio-economical background is important (at least in Germany) in determining students' overall academic course. This factor is beyond the students' control, but motivation and personality – besides cognitive abilities, of course - then seem to “finetune” how successful they are within the academic setting they find themselves in.

It is noteworthy that the items used in this study operationalize their target constructs, such as motivation, on a very global level, without investigating the root causes. We measured whether a child was motivated or not, and what grades it achieved, but do not know which external influences might have led up to this current state. In particular, some of the existing literature on underachievement indicates that peers and teachers can play an important role in shaping a student's educational path (e. g., Baker, 1998; Woodward & Fergusson, 2000; Boehnke, 2008; see also Section 1.3). A striking example comes from Fong and Krause (2014), who found that underachieving college students did not necessarily lack academic self-efficacy, but reported far less encouragement (“social persuasions”) than achieving students. Asbury, Moran and Plomin (2017) identified monozygotic twin pairs whose academic achievement differed and asked them and their parents about possible reasons; different teachers were among the most frequent causal attributions. The current study does not allow conclusions as to why the different groups in our sample differ the way they do and what role teachers and peers play in shaping, or reinforcing, these differences.

As a final point, some criticisms directed at the construct of underachievement, or its interpretation, are worth mentioning. On a very basic level, it should be remembered that underachievers underachieve only relative to their IQ. Stating that they “do not fulfill their potential” assumes that IQ is an all-encompassing, precise measure of a person's potential (Ziegler et al., 2012). Instead, IQ as a predictor of achievement is just that: one predictor. A meteorologist who is surprised by a rainy day after forecasting sunshine will not claim that the day “did not fulfill its potential”, but look for factors that explain the deviation. In the same sense, achievement scores not lining up perfectly with IQ can only come about by a) measurement error or b) other influences on achievement not accounted for. The second possibility is not only more promising, but has also found support through previous research as well as our results. Also note that within the underachievement

paradigm, any factor that prevents a student from doing well in school *and on an IQ test* will, by definition, not be uncovered within the underachievement paradigm.

It has also been pointed out that underachievers do not necessarily have to be unhappy about the situation; they might have decided for themselves that other things than school are more important to them and be fine with that choice (Figg, et al. 2012). Furthermore, the fact that underachievers as a group did not appear to be unusually heterogenous with respect to the variables used here should not be misinterpreted to mean that there is a one-fits-all recipe to help them translate their cognitive abilities into achievement. Individual perspectives are necessary and could be a focus of future research.

5 Conclusions

The present study solidified the empirical basis on which researchers and practitioners alike refer to students with low grades despite high cognitive abilities as a distinct group. That they differ from other highly intelligent students in various personal and environmental aspects was already established; besides replicating these differences, the patterns found in our sample of 51 under- or overachievers and 176 comparison students suggest that underachievers also differ from fellow unsuccessful students in terms of their personality. Motivational variables play a smaller role (if any); while motivation and parental support are lower among underachievers than among their achieving counterparts, this deficit is not extraordinary compared to other low achievers. A particular goal of the current study was to strengthen conclusions about underachieving students as a group, both by defining relatively strict criteria for identifying underachievement in the study sample as well as by comparing them to relevant groups beyond only other gifted students. In this way, we believe our study adds to the existing literature on underachievement by providing a more nuanced by way of its methodology.

In practical terms, this research suggests that while targeting motivation in unsuccessful students is not necessarily wrong, it is not specifically underachievers who are in need of such an intervention. The differences in the Big Five variables do not immediately suggest practical steps, but one point of interest could be the unusually low conscientiousness reported by underachievers; rather than lacking motivation, some of them might rather be in need of being taught skills and strategies to help them organize their academic endeavors effectively.

References

- Baker, J. A., Bridger, R., & Evans, K. (1998). Models of underachievement among gifted preadolescents: The role of personal, family, and school factors. *Gifted Child Quarterly*, 42(1), 5-15. <https://doi.org/10.1177/001698629804200102>
- Balduf, M. (2009). Underachievement among college students. *Journal of advanced academics*, 20(2), 274-294. <https://doi.org/10.1177/2F1932202X0902000204>
- Barbier, K., Donche, V., & Verschueren, K. (2019). Academic (under) achievement of intellectually gifted students in the transition between primary and secondary education: An individual learner perspective. *Frontiers in Psychology*, 10, 2533. <https://doi.org/10.3389/fpsyg.2019.02533>
- Baum, M.A., Klatzka, C.H., Iser, J. & Hahn, E. (2020). TwinLife scales manual: F2F1, CATI 1, & F2F2. *TwinLife Technical Report Series*, 08. ISSN 2512-403X. Bielefeld University / Saarland University.
- Beierlein, C., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2012). Ein Messinstrument zur Erfassung subjektiver Kompetenzerwartungen: Allgemeine Selbstwirksamkeit Kurzsкала (ASKU) [An instrument to measure subjective competency expectations: The general self-efficacy scale (ASKU)]. (GESIS-Working Papers, 2012/17). Mannheim: GESIS - Leibniz-Institut für Sozialwissenschaften.
- Boehnke, K. (2008). Peer pressure: A cause of scholastic underachievement? A cross-cultural study of mathematical achievement among German, Canadian, and Israeli middle school students. *Social Psychology of Education*, 11(2), 149-160. <https://doi.org/10.1007/s11218-007-9041-z>
- Carlin, J. B., Gurrin, L. C., Sterne, J. A., Morley, R., & Dwyer, T. (2005). Regression models for twin studies: a critical review. *International journal of epidemiology*, 34(5), 1089-1099. <https://doi.org/10.1093/ije/dyi153>
- Castejón, J. L., Gilar, R., Veas, A., & Miñano, P. (2016). Differences in learning strategies, goal orientations, and self-concept between overachieving, normal-achieving, and underachieving secondary students. *Frontiers in psychology*, 7, 1438.
- Colangelo, N., Kerr, B., Christensen, P., & Maxey, J. (1993). A comparison of gifted underachievers and gifted high achievers. *Gifted Child Quarterly*, 37(4), 155-160. <https://doi.org/10.3389/fpsyg.2016.01438>
- Cone, T. E., & Wilson, L. R. (1981). Quantifying a severe discrepancy: A critical analysis. *Learning Disability Quarterly*, 4(4), 359-371. <https://doi.org/10.2307/1510737>
- Costa Jr, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and individual differences*, 13(6), 653-665. [https://doi.org/10.1016/0191-8869\(92\)90236-I](https://doi.org/10.1016/0191-8869(92)90236-I)
- Dickhäuser, O., Schöne, C., Spinath, B., & Stiensmeier-Pelster, J. (2002). Die Skalen zum akademischen Selbstkonzept: Konstruktion und Überprüfung eines neuen Instrumentes. [The

Academic Self Concept Scales: Construction and Evaluation of a new Instrument.] *Zeitschrift für differentielle und diagnostische Psychologie: ZDDP*, 23(4), 393-405.

<https://doi.org/10.1024//0170-1789.23.4.393>

Dixon, R. M., Craven, R., & Martin, A. (2006). Underachievement in a whole city cohort of academically gifted children: What does it look like? *Australasian Journal of Gifted Education*, 15(2), 9–15.

Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What No Child Left Behind leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of educational psychology*, 104(2), 439-451.

<https://doi.org/10.1037/a0026280>

Figg, S. D., Rogers, K. B., McCormick, J., & Low, R. (2012). Differentiating low performance of the gifted learner: Achieving, underachieving, and selective consuming students. *Journal of Advanced Academics*, 23(1), 53-71. <https://doi.org/10.1177%2F1932202X11430000>

Gerlitz, J. Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP [On measuring personality using the Big Five in SOEP]. *DIW Research Notes*, 4, 2005.

Goodman, R., Meltzer, H., & Bailey, V. (1998). The Strengths and Difficulties Questionnaire: A pilot study on the validity of the self-report version. *European child & adolescent psychiatry*, 7(3), 125-130. <https://doi.org/10.1007/s007870050057>

Gottschling, J. (2017). Documentation TwinLife Data: Cognitive Abilities. *TwinLife Technical Report Series. Bielefeld/Saarbrücken*. <https://dbk.gesis.org/DBKSearch/download.asp?db=D&id=60848>.

Grözinger, G., & Baillet, F. (2015). Gibt es auch beim Abitur eine Noteninflation? Zur Entwicklung der Abiturnoten als Hochschulzugangsberechtigung-Eine Darstellung und Analyse aus Soziologischer Perspektive. *Bildung und Erziehung*, 68(4), 473. <https://doi.org/10.7788/bue-2015-0407>

Hanses, P., & Rost, D. H. (1998). Das „Drama der hochbegabten Underachiever“: "gewöhnliche" oder "außergewöhnliche" Underachiever? [The "drama" of the gifted underachievers - "ordinary" or "extraordinary" underachievers?] *Zeitschrift für Pädagogische Psychologie*, 12(1), 53-71.

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: causal relationships and underlying mechanisms. *Psychological bulletin*, 111(1), 127.

Hoover-Schultz, B. (2005). Gifted underachievement: Oxymoron or educational enigma? *Gifted Child Today*, 28(2), 46-49. <https://doi.org/10.4219%2Fgct-2005-171>

Hwang, M. H., Lee, D., Lim, H. J., Seon, H. Y., Hutchison, B., & Pope, M. (2014). Academic

- underachievement and recovery: Student perspectives on effective career interventions. *The Career Development Quarterly*, 62(1), 81-94. <https://doi.org/10.1002/j.2161-0045.2014.00072.x>
- Johnson, W., McGue, M., & Iacono, W. G. (2006). Genetic and environmental influences on academic achievement trajectories during adolescence. *Developmental psychology*, 42(3), 514-532. <https://doi.org/10.1037/0012-1649.42.3.514>
- Kuhn, J. T., Holling, H., & Freund, P. A. (2008). Begabungsdagnostik mit dem Grundintelligenztest (CFT 20-R) Psychometrische Eigenschaften und Messäquivalenz. [Assessing giftedness with the CFT 20-R. Psychometric properties and measurement invariance.] *Diagnostica*, 54(4), 184-192. <https://doi.org/10.1026/0012-1924.54.4.184>
- Landis, R. N., & Reschly, A. L. (2013). Reexamining gifted underachievement and dropout through the lens of student engagement. *Journal for the Education of the Gifted*, 36(2), 220-249. <https://doi.org/10.1177/2F0162353213480864>
- Lang, V., & Kottwitz, A. (2017). The sampling design and socio-demographic structure of the first wave of the TwinLife panel study: a comparison with the Microcensus. *TwinLife Technical Report Series* (Vol. 3).
- Lau, K. L., & Chan, D. W. (2001). Identification of underachievers in Hong Kong: Do different methods select different underachievers?. *Educational Studies*, 27(2), 187-200. <https://doi.org/10.1080/03055690120050419>
- Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83(404), 1198-1202.
- Lohmar, B., & Eckhardt, T. (Eds.). (2014). *The Education System in the Federal Republic of Germany 2012/2013*. Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. Retrieved from: http://www.istp2016.org/fileadmin/Redaktion/Dokumente/documentation/dossier_en_ebook.pdf
- McCall, R. B. (1994). Academic underachievers. *Current Directions in Psychological Science*, 3(1), 15-19.
- McCall, R. B., Beach, S. R., & Lau, S. (2000). The nature and correlates of underachievement among elementary schoolchildren in Hong Kong. *Child Development*, 71(3), 785-801. <https://doi.org/10.1111/1467-8624.00185>
- McCoach, D. B., & Siegle, D. (2003). Factors that differentiate underachieving gifted students from high-achieving gifted students. *Gifted child quarterly*, 47(2), 144-154. <https://doi.org/10.1177%2F001698620304700205>
- McGue, M., & Bouchard, T. J. (1984). Adjustment of twin data for the effects of age and sex. *Behavior genetics*, 14(4), 325-343. <https://doi.org/10.1007/BF01080045>

- Mofield, E., & Parker Peters, M. (2019). Understanding Underachievement: Mindset, Perfectionism, and Achievement Attitudes Among Gifted Students. *Journal for the Education of the Gifted*, 42(2), 107-134. <https://doi.org/10.1177%2F0162353219836737>
- Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15), 3301-3307. <https://doi.org/10.1093/bioinformatics/bti499>
- Mönkediek, B., Lang, V., Weigel, L., Baum, M. A., Eifler, E. F., Hahn, E., Hufer, A., Klatzka, C. H., Kottwitz, A., Krell, K., Nikstat, A., Diewald, M., Riemann, R., & Spinath, F. M. (2019). The German Twin Family Panel (TwinLife). *Twin Research and Human Genetics*, 22, 540-547. <https://doi.org/10.1017/thg.2019.63>
- Obergriesser, S., & Stoeger, H. (2015). The role of emotions, motivation, and learning behavior in underachievement and results of an intervention. *High Ability Studies*, 26(1), 167-190. <https://doi.org/10.1080/13598139.2015.1043003>
- O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual differences*, 43(5), 971-990. <https://doi.org/10.1016/j.paid.2007.03.017>
- Peterson, J. S., & Colangelo, N. (1996). Gifted achievers and underachievers: A comparison of patterns found in school files. *Journal of Counseling & Development*, 74(4), 399-407. <https://doi.org/10.1002/j.1556-6676.1996.tb01886.x>
- Plewis, I. (1991). Underachievement: a case of conceptual confusion. *British Educational Research Journal*, 17(4), 377-385. <https://doi.org/10.1080/0141192910170407>
- Preckel, F., Holling, H., & Vock, M. (2006). Academic underachievement: Relationship with cognitive motivation, achievement motivation, and conscientiousness. *Psychology in the Schools*, 43(3), 401-411. <https://doi.org/10.1002/pits.20154>
- Reis, S. M., & McCoach, D. B. (2000). The underachievement of gifted students: What do we know and where do we go? *Gifted child quarterly*, 44(3), 152-170. <https://doi.org/10.1177%2F001698620004400302>
- Ritchotte, J. A., Matthews, M. S., & Flowers, C. P. (2014). The validity of the achievement-orientation model for gifted middle school students: An exploratory study. *Gifted Child Quarterly*, 58(3), 183-198. <https://doi.org/10.1177%2F0016986214534890>
- Rimfeld, K., Kovas, Y., Dale, P. S., & Plomin, R. (2016). True grit and genetics: Predicting academic achievement from personality. *Journal of personality and social psychology*, 111(5), 780-789. <http://dx.doi.org/10.1037/pspp0000089>
- Rost, D. H., & Hanses, P. (1997). Wer nichts leistet, ist nicht begabt? Zur Identifikation

- hochbegabter Underachiever durch Lehrkräfte. [Not achieving ---- not gifted? About the identification of gifted underachievers by teacher-ratings.] *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 29(2), 167-177.
- Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, 53, 118-137.
<https://doi.org/10.1016/j.intell.2015.09.002>
- Rubenstein, L. D., Siegle, D., Reis, S. M., McCoach, D. B., & Burton, M. G. (2012). A complex quest: The development and research of underachievement interventions for gifted students. *Psychology in the Schools*, 49(7), 678-694. <https://doi.org/10.1002/pits.21620>
- Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological bulletin*, 143(6), 565-600.
<https://doi.org/10.1037/bul0000098>
- Siegle, D. (2018). Understanding underachievement. In S. I. Pfeiffer (ed.), *Handbook of Giftedness in Children* (pp. 285-297). Springer, Cham.
- Siegle, D., DaVia Rubenstein, L., & McCoach, D. B. (2020). Do you know what I'm thinking? A comparison of teacher and parent perspectives of underachieving gifted students' attitudes. *Psychology in the Schools*. <https://doi.org/10.1002/pits.22345>
- Siegle, D., McCoach, D. B., & Roberts, A. (2017). Why I believe I achieve determines whether I achieve. *High Ability Studies*, 28(1), 59-72. <https://doi.org/10.1080/13598139.2017.1302873>
- Smith, E. (2007). Considering the experiences of 'underachieving' and 'overachieving' students. *International Journal of Research & Method in Education*, 30(1), 19-32.
<https://doi.org/10.1080/17437270701207702>
- Snyder, K. E., & Linnenbrink-Garcia, L. (2013). A developmental, person-centered approach to exploring multiple motivational pathways in gifted underachievement. *Educational Psychologist*, 48(4), 209-228. <https://doi.org/10.1080/00461520.2013.835597>
- Sparfeldt, J. R., Buch, S. R., & Rost, D. H. (2010). Klassenprimus bei durchschnittlicher Intelligenz – Overachiever auf dem Gymnasium. [Excellent Scholastic Achievement with Average Intelligence – Overachievement in Secondary School.] *Zeitschrift für pädagogische Psychologie*, 24(2), 147-155. <https://doi.org/10.1024/1010-0652/a000012>
- Sparfeldt, J. R., Schilling, S. R., & Rost, D. H. (2006). Hochbegabte Underachiever als Jugendliche und junge Erwachsene: Des Dramas zweiter Akt? [Gifted Underachievers as Adolescents and Young Adults: Second Part of the „Drama of Gifted Underachievers“?] *Zeitschrift für Pädagogische Psychologie*, 20(3), 213-224. <https://doi.org/10.1024/1010-0652.20.3.213>
- Spinath, B., Stiensmeier-Pelster, J., Schöne, C., & Dickhäuser, O. (2002). *Skalen zur Erfassung der*

- Lern-und Leistungsmotivation: SELMO* [Scales for the measurement of learning motivation and achievement motivation: SELMO]. Göttingen: Hogrefe.
- Spinath, F. M., & Wolf, H. (2006). CoSMoS and TwinPaW: Initial report on two new German twin studies. *Twin Research and Human Genetics*, 9(6), 787-790. <https://doi.org/10.1375/twin.9.6.787>
- Stankov, L., & Lee, J. (2014). Quest for the best non-cognitive predictor of academic achievement. *Educational Psychology*, 34(1), 1-8. <https://doi.org/10.1080/01443410.2013.858908>
- Steenbergen-Hu, S., Olszewski-Kubilius, P., & Calvert, E. (2020). The effectiveness of current interventions to reverse the underachievement of gifted students: Findings of a meta-analysis and systematic review. *Gifted Child Quarterly*, 64(2), 132-165. <https://doi.org/10.1177/0016986220908601>
- Sternberg, R. J., Grigorenko, E. L., & Bundy, D. A. (2001). The predictive value of IQ. *Merrill-Palmer Quarterly* (1982-), 1-41.
- Tabachnick, B. G., & Fidell, L. S. (2001). Cleaning up your act: Screening data prior to analysis. *Using multivariate statistics*, 5, 61-116.
- Uhlig, J., Solga, H., & Schupp, J. (2009). Bildungsungleichheiten und blockierte Lernpotenziale: Welche Bedeutung hat die Persönlichkeitsstruktur für diesen Zusammenhang? [Inequality in Educational Opportunities: Underachievement and the Role of Personality Traits.] *Zeitschrift für Soziologie*, 38(5), 418-440. <https://doi.org/10.1515/zfsoz-2009-0505>
- Uno, M., Mortimer, J. T., Kim, M., & Vuolo, M. (2010). "Holding on" or "coming to terms" with educational underachievement: A longitudinal study of ambition and attainment. In: S. Shulman & J.-E. Nurmi (Eds.), *The role of goals in navigating individual lives during emerging adulthood. New Directions for Child and Adolescent Development*, 130, 41–56. <https://doi.org/10.1002/cd.280>
- Veas, A., Castejón, J. L., O'Reilly, C., & Ziegler, A. (2018). Mediation analysis of the relationship between educational capital, learning capital, and underachievement among gifted secondary school students. *Journal for the Education of the Gifted*, 41(4), 369-385. <https://doi.org/10.1177/00162353218799436>
- Vlahovic-Stetic, V., Vidovic, V. V., & Arambasic, L. (1999). Motivational characteristics in mathematical achievement: A study of gifted high-achieving, gifted underachieving and non-gifted pupils. *High Ability Studies*, 10(1), 37-49. <https://doi.org/10.1080/1359813990100104>
- Weiß, R., & Osterland, J. (2012). *Grundintelligenztest Skala 1* (CFT 1, 6. überarb. Aufl.). Göttingen: Hogrefe.
- White, S. L., Graham, L. J., & Blaas, S. (2018). Why do we know so little about the factors associated with gifted underachievement? A systematic literature review. *Educational Research*

Review, 24, 55-66. <https://doi.org/10.1016/j.edurev.2018.03.001>

Woodward, L. J., & Fergusson, D. M. (2000). Childhood peer relationship problems and later risks of educational under-achievement and unemployment. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 41(2), 191-201. <https://doi.org/10.1017/S002196309900520X>

Ziegler, A., Ziegler, A., & Stoeger, H. (2012). Shortcomings of the IQ-based construct of underachievement. *Roeper Review*, 34(2), 123-132.
<https://doi.org/10.1080/02783193.2012.660726>

Sports club participation impacts life satisfaction in adolescence: a twin study

Alexander Dings*, Frank M. Spinath

Universität des Saarlandes, Saarbrücken, Germany

* Corresponding author.

E-mail address: alexander.dings@uni-saarland.de (A. Dings).

Abstract

Using data from the German TwinLife panel, we investigate both between-pair and within-pair associations of frequent participation in a sports club with life satisfaction among adolescents and young adults. We employ regression models suitable for twin data to investigate whether presumable links between sports club participation and life satisfaction can be explained through confounding via shared genetic or environmental factors or whether they, in contrast, remain significant within monozygotic (MZ) twin pairs. Results show that participation is associated with higher general as well as domain-specific satisfaction, but that these links differ in their etiology. Genetic and shared environmental confounding fully explains the effects on general life satisfaction, with MZ twins discordant for sports participation not differing on this variable. Conversely, for domain-specific satisfaction, especially related to subjective health, significant within-pair effects among MZ twins point to environmentally mediated effects, consistent with possible causal links. However, tentative longitudinal analyses suggest that these effects might be confounded by previous levels of life satisfaction. Overall, the present results suggest that participation could causally impact domain-specific satisfaction more so than general life satisfaction, but further research is necessary in this domain.

1 Background

1.1 Introduction

A large amount of research has investigated precursors to subjective well-being in general and high life satisfaction in particular, in other words, what makes people happy. Some of the more commonly cited antecedents of well-being and life satisfaction are fulfilling relationships, personality factors such as low neuroticism, a sense of purpose and meaning, and a healthy lifestyle, including sports, exercise or physical activity in general (for reviews, see Diener et al., 2018; Diener & Ryan, 2009; Veenhoven, 1996; Proctor et al., 2009).

However, as in many other research domains, direction and extent of causality are less clearly established (Diener et al., 2018; Veenhoven, 1996). One approach for supporting hypotheses of causality is to use genetically informed samples, such as twin pairs (Pingault et al., 2018; McAdams et al., 2021). Direct evidence of causality, with no room for alternative explanations, remains elusive under this framework as well. Twin models can, however, cover some of the distance towards this ideal. For variables that cannot be experimentally manipulated, monozygotic (MZ) twin pairs that happen to differ (be discordant) on this variable provide a decent approximation to two experimental groups where subjects in one condition ideally do not systematically differ from the other group in variables other than the manipulated variable (McGue et al., 2010). Thus, differences within MZ twin pairs are at the core of using twin samples to strengthen causal inferences.

The present study uses a twin sample to investigate possible associations between life satisfaction and one of its most frequently cited possible causes – physical activity, and participation in sports clubs in particular.

1.2 Physical activity and life satisfaction

Numerous studies show associations between physical activity and well-being (e. g., Buecker et al., 2021; Wiese et al., 2018; Cerin et al., 2009). Among others, this was shown in a large sample taken from the German Family Panel (Schmiedeberg & Schröder, 2016) and in longitudinal analyses that point to possible causal links (Varca et al., 1984). Physical activity is seen as a major component of life satisfaction among adolescents (Proctor et al., 2009) and appears to be effective in reducing symptoms of depression and other mental illnesses (Fox, 1999). Research during the COVID-19 pandemic identified physical activity as a protective factor for mental health (Pietsch et al., 2022). It is less clear *which* kinds of physical activity, and in which context, are beneficial. For instance, vigorous (as opposed to moderate) training may backlash, with one study finding it associated with *lower* well-being (Wicker & Frick, 2015). Evidence on the associations between usual everyday

activity, as opposed to organized exercise, and life satisfaction is generally positive, but mixed (for an overview, see Maher et al., 2017), with some studies finding no effect (e. g., Pedišić et al., 2015). In contrast, there is some evidence that participation in sports clubs not only is generally associated with better physical and mental health (Larsen et al. 2021), but also that such effects are more pronounced than for other kinds of physical activity. Participation in sports clubs was associated with good health to a greater extent than leisure time physical activity in general in a German sample (Spengler & Woll, 2013), and studies by Mutz et al. (2021) as well as Vilhjalmsson and Thorlindsson (1992) show more pronounced effects on general and domain-specific well-being for sports club participation than for other forms of exercise. As social connectedness is often found to be linked to well-being and life satisfaction (e. g., Becchetti et al., 2012; Jose et al., 2012; O'Rourke et al., 2017), the “club” aspect of sports clubs may be as pivotal as the “sports” aspect in promoting well-being. For example, sports clubs may foster social integration (e. g., Elmoose-Østerlund et al., 2019) and participation in them is associated with higher social functioning (Eime et al., 2010).

1.3 A behavioral genetics perspective on physical activity and life satisfaction

In the classical twin design, variance in a trait is decomposed variance into additive (A) genetic influences as well as shared (C) and nonshared (E) environmental effects Depending on the specific approach, non-additive genetic influences (particularly dominance effects, labeled D) can also be incorporated in addition to, or more typically instead of, A. For an overview on this and other analytical approaches to twin data, see Rijdsdijk & Sham (2002). It is important to note that estimates of heritability and the other components are sample-specific and do not represent universally “true” values (see Plomin et al., 2016).

With regards to physical activity, the extent of heritability partially appears to depend on what kind of activity is measured. Objective measurements of how much young children move during the day are mostly attributable to shared environmental influences, less to heritable ones (Fisher et al., 2010). On the other hand, twin studies with adolescents and adults generally show that the differences between individuals in leisure-time physical activity are partly explained by their different genetic background, i. e. heritable (Carlsson et al., 2006; Mustelin et al., 2012). In line with the aforementioned context-dependent nature of heritability estimates, some studies have found moderation effects: heritable influences are especially pronounced in young compared to older adults (Vink et al., 2011) and at higher compared to lower levels of activity (Duncan et al., 2008).

Twin models have been used to investigate effects of physical activities through environmental pathways, i. e. by exploiting MZ twin differences as described above. One finding suggests that

physical activity lowers mortality risk, although the study in question did not include enough MZ twin pairs discordant for activity to yield robust conclusions (Kujala et al., 1998). Leskinen et al. (2009) showed physical benefits (lower high-risk body fat) for the active twins in discordant twin pairs. Korhonen et al. (2009) showed that physical activity was a protective factor against future illicit drug use. BMI is less heritable among physically active twins, suggesting an environmentally mediated link (Mustelin et al., 2009). Tentative results also point to higher average life satisfaction for the physically active twin in a small number of discordant MZ pairs (Waller et al., 2010). General life satisfaction is generally found to be moderately heritable, although estimates vary considerably (for an overview, see Bartels, 2015). The remaining variance appears to be accounted for by unshared environmental influences, with shared environment only being a significant factor in relatively few studies (e. g., Nes et al., 2008). Hahn et al. (2013) found that the heritable proportion of the variance overlaps with genetic sources of variance in personality, and low neuroticism and high extraversion in particular. The remaining variance in Hahn et al.'s (2013) study, however, was explained by nonshared environmental influences which showed no relations to personality.

1.4 The elusive nature of specific nonshared environmental influences

The findings cited above, pointing to a much more prominent role of nonshared rather than shared environment in the etiology of life satisfaction, are similar to what research in behavioral genetics has shown in many areas of personality, health and development. It is in fact such a common pattern across psychology that Turkheimer (2000) postulated a “substantial portion of the variation in complex human behavioral traits” (p. 160) not being explained by genetic or shared environmental effects as his third of three “laws of behavior genetics”. More recently, Plomin et al. (2016) reiterated this point as one of ten reliably replicated findings from behavioral genetics. An extensive meta-analysis of virtually all classical twin studies published at the time showed that nonshared environmental factors are more impactful than shared factors in the vast majority of traits and studies, including many cases where the best-fitting model indicates no influence of shared environmental factors at all (Polderman et al., 2015).

In explaining why nonshared environmental influences are so important across a wide range of domains, two things must be noted. First, at a statistical level, the nonshared environmental component necessarily includes measurement error (Rijsdijk & Sham, 2002). Furthermore, it is important to note that “nonshared” environment, as a variance components in twin models, does not only include external influences that are *objectively* different between twins. Instead, the terms shared and nonshared, in this context, refer to the similar or dissimilar *effect* of an external influence

on different children (Smith, 2011). The distinction is sometimes referred to as objective versus effective environments (e. g., Turkheimer & Waldron, 2000). A shared variable in the objective environment, such as socio-economic status, may be a nonshared variable in the effective environment by impacting siblings differently. On the other hand, the reverse is not possible: objectively nonshared factors, such as having different teachers in school, cannot systematically act in a way that would make siblings and twins more alike, i. e. act as an effectively shared environmental factor. The nonshared environmental component therefore includes all objectively nonshared factors *plus* some objectively shared factors. As highlighted by Turkheimer and Waldron (2000), in this light, the finding that nonshared environment influences explain substantial proportions of variance in most traits is less surprising.

However, pinpointing specific sources of nonshared environmental variance has proven to be the proverbial search for a needle in a haystack. Spanning several decades, Plomin and co-authors have repeatedly discussed the empirical and theoretical state of nonshared environment in behavior genetics (Rowe & Plomin, 1981; Plomin & Daniels, 1987; Plomin et al., 2001; Plomin, 2011). Across these articles, and in still more recent work as well (e. g., Gidziela et al., 2023), the respective authors stress that identifying specific influences which act to siblings different from one another has been difficult. For instance, Plomin (2011) noted that “the big question for nonshared environment is how to identify the ‘missing’ nonshared environment” (p. 585). This is reminiscent of the “missing heritability” term coined to describe the (initially) tedious search for specific genotypes that make up heritable sources of variance (Maher, 2008). Earlier, Plomin et al. (2001) had considered an even more pessimistic outlook, which they termed the “gloomy prospect” (p. 231). This refers to the view that broad outcomes such as happiness might be shaped through an accumulation of tiny effects, which furthermore might be highly idiosyncratic in nature and therefore difficult to detect as main effects across a population. In line with this notion, empirical attempts to identify concrete factors that can explain the large nonshared environmental components of variance were met with little success (Turkheimer & Waldron, 2000). Other authors have since argued that a large part of nonshared environmental influences might in fact reflect unsystematic, idiosyncratic factors that are indistinguishable from random chance events (Smith, 2011), including stochastic fluctuations at the molecular genetic level (Tikhodeyev & Shcherbakova, 2019).

1.5 Establishing environmentally mediated links through twin differences

Most studies that attempt to identify nonshared environmental effects on behavior make use of twin / co-twin differences, especially among MZ twin pairs. The underlying logic is that since MZ twins are virtually genetically identical, any measured difference between the two members of an

MZ twin pair can by definition not be explained by genetic sources of variance (commonly denoted by A in twin models). Similarly, any non-genetic influences that affect them in the same way (shared environment - C) can, by definition, also be ruled out as an explanation (Pike et al., 1996). Therefore, using MZ differences in a regression approach is equivalent to a regression where *every* variable that affects the outcome through genetic or shared environmental pathways has been entered as a control variable. The truism that “correlation is not causation” still applies, especially because confounding can also arise from nonshared factors if they are not explicitly controlled for (Sjölander et al., 2012). Still, this method rules out a wide range of alternative explanations implicitly, without the need for direct measurement. For a more detailed overview on this approach, we refer readers to Vitaro et al. (2009).

In accordance with the pessimism expressed in Plomin et al.'s (2001) “gloomy prospect”, such research often finds a seemingly external influence on an outcome to actually be fully confounded by shared factors – common genetic (or shared environmental) factors act as a third variable that impacts both traits and thereby creates an illusory correlation. For example, this approach has been used to show that while negative life events are generally associated with lower quality of sleep, this seemingly obvious link disappears when the analysis focuses on differences within MZ pairs, thereby controlling for common genetic background and shared environmental variance (Barclay et al., 2012).

In the same vein, some ostensible influences on life satisfaction appear to be spurious when viewed through a behavioral genetic lens. For example, in accordance with the results cited above (Hahn et al., 2013), being extroverted and low in neuroticism is associated with higher life satisfaction in the general population. However, this is not the case when analyzing MZ twin differences (Misheva, 2016). Similarly, it seems almost trivial to assume that harsh parental discipline directly lowers life satisfaction, or that being unemployed impacts it causally, but these associations vanished in MZ difference analyses (see Mann et al., 2019, for parental discipline; Schnittker, 2008, for employment status).

Arguably the most striking finding in this regard comes from a study by Stubbe et al. (2007), which compared twins who regularly engage in exercise to their co-twins who do not. Within monozygotic (MZ) pairs, there was no significant difference in happiness between such discordant twins, suggesting that the overall correlation between the two variables is not causal, but instead the result of genetic (or shared environmental) confounding, as described above. If the same genetic characteristics predispose people both to engaging in exercise and to being happy, these two outcomes will be correlated even if one does not cause the other, and the finding by Stubbe et al. (2007) suggests that this might be the case.

Considering the large number of studies that link physical activity to well-being, evidence that this correlation might be spurious is surprising and would constitute an impactful finding for the field if backed up by further studies. As such, these results invite attempts to replicate them in other samples. It is worth noting that some precursors to life satisfaction do persist in MZ twin difference analyses. In particular, friend support remained significant a significant predictor of well-being both in Mann et al. (2019) and Schnittker (2008), strengthening the evidence for the importance of social connectedness. Overall, success in identifying specific nonshared environmental influences on life satisfaction is mixed, and investigations into the role of physical activity under this framework are scarce and inconclusive.

2 Hypotheses

The present study uses data from a twin panel study with the aim of clarifying possible unmeasured familial confounding of presumed links between adolescents' sports club participation and life satisfaction. Although there is no shortage of findings that link physical activity – and participation in sports clubs in particular - to well-being, genetically informed evidence is, to our knowledge, limited to two studies with conflicting results (Waller et al., 2010; Stubbe et al., 2007). The study aims to contribute to this existing mixed evidence.

More so than in other countries, local associations based on shared pastimes (*Vereine*) are a prominent part of social life in German culture, with about six associations per 1,000 inhabitants (Franzen & Botzen, 2011). Sports clubs are the most frequent kind among them, playing a major role in the country's economy (Breuer et al., 2015a). In contrast to the situation in the United States, where most high school students participate in an athletic team within the school (Veliz et al., 2019), extracurricular activity plays a relatively small role in Germany, as sports clubs are typically independent from schools (Schmidt et al., 2020). We hypothesize that regularly participating in such a club should phenotypically be correlated with life satisfaction because it arguably combines two variables commonly implicated in life satisfaction: physical activity and social connectedness. We then use twin regression models to investigate whether such correlations, if they are present in the current sample, can be explained through genetic confounding or remain significant within MZ twin pairs. As previous research is inconclusive and the present study does not build on a specific theory that would suggest either the presence or absence of such an effect beyond confounding, we regard this analysis as exploratory.

3 Methods

3.1 Sample

The present study uses data from TwinLife, an ongoing twin panel study of about 4000 twin pairs, and their families, in Germany starting in 2014 (Mönkediek et al., 2019; for details on the sampling procedure and characteristics of the final sample, see Lang & Kottwitz, 2020). This panel includes four birth cohorts of twins. At the time of the first interview, twins in these cohorts were approximately 5, 11, 17 and 23-24 years old, respectively. Data from TwinLife are publicly available via the GESIS data archive (<https://doi.org/1.4232/1.14118>).

3.2 Measures

Information on sports club participation was available for the three older cohorts out of the four in the first data collection wave. For cohort 2 (about 11 years old at the time), parents answered for their kids. This cohort was not included in the final sample for the analyses because it included very few discordant MZ pairs (see below). For older cohorts, participation was self-reported by the twins.

Among the four response categories on sports club participation, “never” (55.6%) and “weekly” (36.9%) were far more prevalent than the intermediate categories (“every month” at 4.9% and “less than every month” at 2.6%). We grouped these intermediate cases together with those who reported they never went to a sports club, i. e. dummy-coded participation as 1 (weekly) or 0 (rarely or never). This was done with the goal of enabling a more straightforward interpretation of the results, as 1 now stands for frequent participation in a sports club and 0 for no such (regular) participation. While dichotomizing data is generally not recommended (Irwin & McClelland, 2003), it is more justifiable when the data clearly do not follow a normal distribution and the underlying variable may more sensibly be interpreted as an ordinal rather than an interval scale (Streiner, 2002). As the interpretation of the response categories on the current question as an interval scale appears doubtful, the data are clearly not normally distributed and the discarded categories only included 7.5% of participants, we concluded that the benefit of easier interpretation of results outweighed the small loss of information.¹⁰

Since the analyses fundamentally rely on differences within MZ pairs, those differences must occur frequently enough in the sample to provide a decent sample size. This is not the case for club participation among the youngest cohort of twins, where only 11 MZ pairs out of 302 were discordant on this (dichotomized) variable (see Results). Notably, almost 90% of dizygotic (DZ)

¹⁰ As a sensitivity check, results of the subsequent analyses do not differ meaningfully when this variable is not dichotomized.

pairs in this age group were concordant as well. A pattern of high similarity within MZ twin pairs, but almost equal similarity within DZ twin pairs indicates that neither nonshared environmental nor genetic factors play a large role (see Rijdsdijk & Sham, 2002). It follows that the remaining source of variance – shared environmental factors – is of high importance. A possible explanation for this pattern could be that presumably, joining a sports club at this age may typically be a decision made by the parents jointly for both kids, rather than the kids themselves individually. As the analyses in the current studies rely heavily on differences between twins, and MZ twins especially, this youngest cohort, with only 11 discordant MZ pairs, was excluded from subsequent analyses. In older cohorts, discordant pairs were more common (and participation became less common in general), and concordance rates among MZ twins differed more notably from those among DZ twins. Since the difference between concordance rates or correlations in MZ versus in DZ pairs provides an index of heritability, this suggests that heritable sources of variance explain an increasing amount of variance in sports club participation as children grow into adolescents and young adults.

General life satisfaction was measured with the *Satisfaction with Life Scale* (SWLS; Diener et al., 1985). Cronbach's alpha for this scale, which consists of 5 items answered on a scale of 1 to 5, was $\alpha = .84$ in the current sample, indicating excellent internal consistency. Domain-specific life satisfaction was measured using ad-hoc items that asked participants: “How satisfied are you...” and providing an 11-point response scale from “not at all satisfied” (0) to “completely satisfied” (10) to rate satisfaction in several domains. The items used in the present study were: “...with your health”, “with your leisure time” and “with your circle of friends and acquaintances”. Other items (satisfaction with romantic relationships, family life, school, income and occupation) were not included because these did not have any straightforward connection to the benefits plausibly associated with activity in a sports club and / or because they did not apply to all participants.

3.3 Analysis

As a precursor to the main analyses, the Results section includes intraclass correlations (ICCs) among MZ and DZ twins as well as parameters estimated from structural equation models to estimate the contribution of genetic and environmental factors to the study variables. Intraclass correlations, in this context, refer to the phenotypic correlations within twin pairs, presented separately for MZ and DZ twin pairs. They can be used to estimate the heritability, or h^2 , of a trait via Falconer's formula (Falconer, 1965), where $h^2 = 2*(r_{MZ} - r_{DZ})$. By extension, the influence of shared environmental factors can be inferred as $c^2 = r_{MZ} - h^2$ (Rijdsdijk & Sham, 2002). The remaining variance ($1 - h^2 - c^2$) can be attributed to nonshared environmental factors (referred to as

e^2). These calculations are valid only under certain conditions, particularly *proportion homogeneity* (i. e., the proportions of a^2 , c^2 and e^2 are assumed to be the same in MZ and DZ twins; see Arbet et al., 2020). However, these formulas only serve as approximations, and structural equation models using maximum likelihood criteria can more accurately decompose variance into additive (A) and non-additive (D, for dominance effects) genetic influences as well as shared (C) and nonshared (E) environmental effects. Because C and D are perfectly confounded in twin pairs reared together, classic twin models cannot estimate them if only data from MZ and DZ twins is available (see Røysamb & Tambs, 2016). Instead, fit statistics are used to compare a model including A, C and E to an ADE model. In addition, if any of the variance components are close to zero, a reduced, more parsimonious model might be preferable. In the present study, these analyses were conducted using the R package *umx* (Bates et al., 2019). The *umxReduceACE* function in this package was utilized to select the best-fitting model according to Akaike's Information Criterion (AIC), which balances model fit and parsimony (Akaike, 1987). In general, all analyses in the present study were performed using R Statistical Software (v4.1.2; R Core Team, 2021).

As described earlier (see section 1.5), linear regressions that use twin difference data are helpful for untangling effects of nonshared environmental factors from those explained by genetic and environmental factors common to both twins in a pair. Historically, such regression approaches to analyzing twin data have typically focused on the differences within MZ pairs only (see Carlin et al., 2005). More recently, a “mixed strategy” has become more common. Besides within-pair differences, this approach also takes into account between-pair differences, that is, a twin pair's mean value in the variable of interest compared to the other pairs in the sample. This can be expressed by the following regression equation:

$$E(Y_{ij}) = \beta_0 + \beta_W * (X_{ij} - \bar{X}_i) + \beta_B * \bar{X}_i$$

Here, the between-pair factor (β_B) measures effects resulting from factors common to both twins in a pair, i. e., shared genetic or environmental background. In statistical terms, “it gives the expected change in Y for a one-unit change in the twin-pair average X, while holding the individual deviation from the average constant” (Carlin et al., 2005, p. 1092). In contrast, the within-pair factor (β_W) indicates changes in the outcome that are associated with each individual twin's deviation from the pair's average (i. e., differences between a twin and their co-twin) while controlling for the pair's mean value. Significant values for β_W therefore indicate that the observed effects cannot be explained by genetic and environmental confounding alone, instead pointing to nonshared environmental effects. If, in fact, *only* the β_W coefficient is significant, confounding by shared

factors appears to play no role at all. Conversely, a significant value for β_B while β_W does not significantly differ from zero can be taken as evidence that such confounding factors can *fully* explain the observed effects. In this case, the twins in an MZ pair will on average not differ from one another in Y no matter how much they differ in X. If between-pair and within-pair effects are both significant, results should accordingly be interpreted to reflect both mechanisms, taking into account their relative magnitudes. For further details on the mixed strategy and its interpretation, we refer readers to Carlin et al. (2005).

We use a linear mixed regression model (not to be confused with the “mixed strategy” outlined above, which refers to the separation of within-pair and between-pair effects) with the index of a twin pair in the data set modeled as random effects to account for the fact that twin data is clustered in pairs (see Carlin et al., 2005). Age and sex, which are necessarily shared by twins (the TwinLife panel only includes same-sex DZ pairs), were included as covariates. All variables were z-standardized, with the exception of sports club participation. As this is a binary “yes / no” variable (see Results section), its unstandardized values of 1 and 0 have a straightforward interpretation. We follow the procedure in Arsenaault et al. (2008) in reporting the regression results by first including the full sample (that is, including DZ twins), followed by only the MZ twins and as a third step using future (i. e., measured at the second measurement wave in TwinLife) rather than concurrent life satisfaction as the outcome, controlling for current levels of life satisfaction. DZ twin pairs control for confounding through shared environment to the same extent as MZ twins, assuming that MZ twins are not systematically treated more similarly than DZ twins (empirical tests of this so-called *equal environments assumption* typically confirm it or find only small biases; see Felson, 2014). However, since DZ twins are no more similar genetically than regular siblings, sharing 50% of their segregating genes on average, genetic similarities between the twins will be less pronounced in this full sample than among MZ twins only. Therefore, DZ twins might differ in genetic factors that impact both sports club participation and life satisfaction, i. e. display genetic confounding. Since this is not possible among MZ twins, a decrease of the between-pair coefficient β_W in the MZ pairs sample compared to the sample of MZ *and* DZ twins implies that these between-pair effects are at least partly driven by genetic confounding (as opposed to shared environmental confounding, which is as controlled for among DZ twins as among MZ twins). We apply these analyses to both general and domain-specific life satisfaction as these different conceptualizations of well-being offer distinct value (Long & Huebner, 2014) with only moderate overlap (Lent et al., 2005). If participating in a sports club is indeed beneficial for adolescents' well-being, the domain-specific items may help clarify why that is the case.

To further illustrate the results of this analysis, we also use an approach exemplified in Stubbe et al. (2007). The dichotomous nature of the predictor variable, which leads to twin pairs either being concordant or discordant for sports participation, can be used to illustrate the pattern behind associations that are confounded through common genetic or environmental factors versus those that are not. If participation causally affects life satisfaction, twin pairs where only one twin participates in a sports club should differ in this regard, and crucially, this should hold true not only for DZ twins, but also for MZ twins. If an individual's genetic makeup explains both their levels of sports club participation and of life satisfaction, leading to a spurious correlation, MZ twins who do participate in a sports club should not, on average, be happier than their genetically identical twins.

4 Results

The full sample used for the analysis included 3492 twins from cohorts 3 (around 17 years old) and 4 (around 23-24 years old), 58% of which were female. These included 423 MZ and 483 DZ pairs from cohort 3 and 454 MZ and 386 DZ pairs from cohort 4. More detailed summary statistics grouped by cohort, zygosity and gender are presented in the Supplementary Materials.

Means and standard deviations for the variables used, as well as their intercorrelations, are provided in Table 1.

Table 2 shows concordance rates for the dichotomous variable of frequent sports club participation by cohort. Similarly, Table 3 displays intraclass correlations (ICCs) for the continuous variables. Prior to calculating ICCs, each variable was regressed on age and sex to control for the paired nature of twin data (see McGue & Bouchard, 1984).

While there is no consistent pattern in the ICCs for the various measures of general and domain-specific life satisfaction in the current sample, MZ correlations in Table 3 generally exceed DZ correlations to a small or moderate degree, suggesting heritabilities between .00 (for domain-specific satisfaction with leisure activities at time 1) and .52 (domain-specific satisfaction with leisure activities at time 1).

The results from the structural equation models presented in Table 4 are mostly consistent with previous findings that nonshared environmental sources of variance play a large role¹¹ and heritability is modest. For domain-specific life satisfaction with health and leisure time activities, modeling the genetic effects as dominance effects (i. e., non-additive effects) resulted in a better fit than the standard ACE model.

11 Nonshared environmental variance includes measurement error by default, which may be more substantial than usual for the domain-specific satisfaction categories, as these were measured by single items.

Table 1*Descriptive statistics and correlations for study variables.*

Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9	10	11
1. age	19.9	3.07	-										
2. sex ^a	.58	-	.02	-									
3. GLS (time 1)	3.86	.79	-.03	.00	-								
4. DLS: health (time 1)	7.82	2.19	-.08	-.04	-.31	-							
5. DLS: leisure (time 1)	7.52	2.19	-.07	-.04	-.31	-.36	-						
6. DLS: friends (time 1)	8.44	1.89	-.06	-.00	-.26	-.39	-.49	-					
7. GLS (time 2)	3.84	.79	-.01	-.00	-.63	-.26	-.22	-.22	-				
8. DLS: health (time 2)	7.55	2.24	-.04	-.04	-.31	-.44	-.17	-.15	-.37	-			
9. DLS: leisure (time 2)	7.30	2.1	-.07	-.05	-.29	-.18	-.33	-.24	-.35	-.29	-		
10. DLS: friends (time 2)	8.25	1.78	-.04	-.01	-.24	-.12	-.19	-.32	-.32	-.27	-.45	-	
11. sports club participation ^b	.47	-	-.16	-.10	-.14	-.09	-.12	-.08	-.13	-.09	-.10	-.09	-

Note: ^a 0 = male, 1 = female. ^b 0 = no frequent participation (includes “never”, “less than once a month” and “monthly”; see section 3.2) and 1 = weekly participation. GLS = general life satisfaction (range: 1 - 5), DLS = domain-specific life satisfaction (range: 1 - 10). *n* = 3492 for all variables. **p* < .05. ***p* < .01.

Table 2*Concordance rates for weekly sports club participation by cohort and zygosity.*

Cohort	Zygosity	Weekly sports club participation		
		Both twins	One twin	Neither twin
2 (around 11 years old)	MZ	204 (67.5%)	11 (3.6%)	87 (28.8%)
	DZ	304 (67.1%)	51 (11.3%)	98 (21.6%)
3 (around 17 years old)	MZ	202 (47.8%)	61 (14.4%)	160 (37.8%)
	DZ	193 (4.0%)	136 (28.2%)	154 (31.9%)
4 (around 23 years old)	MZ	141 (31.1%)	90 (19.8%)	223 (49.1%)
	DZ	71 (18.4%)	128 (33.2%)	187 (48.4%)

Note: Because of the low proportion of discordant MZ pairs in cohort 2, only cohorts 3 and 4 were used in the final analysis (see section 3.2).

Table 3*Intraclass correlations for measure of life satisfaction.*

Variable	Zygosity	
	MZ	DZ
GLS (time 1)	.39 [.34, .45]	.31 [.25, .37]
DLS: health (time 1)	.23 [.15, .31]	.05 [-.04, .13]
DLS: leisure (time 1)	.15 [.07, .23]	.15 [.07, .23]
DLS: friends (time 1)	.32 [.23, .41]	.06 [-.04, .16]
GLS (time 2)	.40 [.31, .48]	.33 [.24, .42]
DLS: health (time 2)	.29 [.17, .39]	.12 [.00, .23]
DLS: leisure (time 2)	.27 [.16, .36]	.20 [.09, .30]
DLS: friends (time 2)	.22 [.09, .34]	.06 [-.06, .19]

Note: GLS = general life satisfaction, DLS = domain-specific life satisfaction. n = 877 MZ pairs and 869 DZ pairs. Confidence intervals for ICCs presented in brackets.

Table 4*Estimates from best-fitting twin models for measures of life satisfaction.*

	A	C	E	D
GLS (time 1)	.21 [.08, .36]	.20 [.08, .32]	.59 [.54, .65]	-
DLS: health (time 1)	-	-	.72 [.66, .79]	.28 [.21, .34]
DLS: leisure (time 1)	.22 [.16, .27]	-	.78 [.73, .84]	-
DLS: friends (time 1)	-	-	.69 [.63, .75]	.31 [.25, .37]
GLS (time 2)	-	.35 [.29, .41]	.65 [.59, .71]	-
DLS: health (time 2)	-	-	.84 [.77, .90]	.16 [.14, .23]
DLS: leisure (time 2)	-	.16 [.14, .23]	.84 [.77, .90]	-
DLS: friends (time 2)	-	-	.78 [.70, .87]	.22 [.13, .30]

Note: Parameter estimates from structural equation models for measures of life satisfaction in the full sample (both cohorts; n = 877 MZ pairs and 869 DZ pairs). A = additive genetic effects, C = shared environmental effects, E = nonshared environmental effects, D = dominance effects (non-additive genetic effects). GLS = general life satisfaction, DLS = domain-specific life satisfaction. For each variable, only the best-fitting model according to the Akaike Information Criterion (AIC) is shown. Confidence intervals for parameter estimates presented in brackets.

Both the general life satisfaction scale as well as the domain-specific satisfaction items were somewhat negatively skewed (skew values between -.72 and -1.99). Still, raw data were used to estimate linear mixed models as transformations can introduce bias in regression models (Schmidt & Finan, 2018) and are not typically necessary in linear mixed models, especially in large samples

such as the current one (Schielzeth et al., 2020)¹². All regression models include age and sex as covariates. Table 5 shows the regression coefficients as well as confidence intervals¹³.

Results show that sports club participation is associated with higher life satisfaction at the general and domain-specific level, but that patterns of within-pair and between-pair effects differ. For general life satisfaction, within-pair effects are evident only in the full sample (i. e., when DZ pairs are included). Among MZ pairs, between-pair effects remain, but within-pair effects do not. However, within-pair effects stay significant for the links between sports club membership and domain-specific happiness with health and leisure activities; for domain-specific satisfaction with one's social life, the within-pair effect is approximately as large within MZ pairs as in the full sample, but does not reach statistical significance in the smaller MZ subset (neither does the between-pair effect).

All of these associations disappear when predicting future satisfaction beyond the current baseline. Here, both within-pair and between-pair coefficients for current happiness are significant predictors, indicating that stability in life satisfaction is driven both by variance components common to an MZ twin pair and those that differ between them.

As an additional illustration of these results, Table 6 shows that for general life satisfaction, only DZ twins discordant for sports club participation, but not discordant MZ twins, tend to differ in life satisfaction. For the domain-specific item of satisfaction with health, however, a significant difference remains even within MZ-pairs. Similarly, for satisfaction with leisure, twins engaging in sports clubs appear to be consistently happier on average, both in DZ and MZ pairs, although statistical significance at $p < .05$ is only marginally achieved in the MZ group. The sample size for the present analysis, which necessarily focuses on discordant pairs (with fewer instances among MZ than DZ twins) may not have been quite sufficient to test differences of about 0.2 – 0.3 standard deviations with a high degree of confidence. Of note, for health and leisure time satisfaction, the numerical differences among the MZ twins are in fact greater than those among the DZ twins.

12 Running a generalized mixed regression model with log-transformed data (and a Gaussian distribution) instead yields the same conclusions. We present the linear mixed model with raw data for ease of interpretation.

13 While we denote significant values as such in the table, inference statistics in mixed regression models must be interpreted with caution; we report them as provided in the R package *parameters* (Lüdtke et al., 2020).

Table 5

Coefficients from linear mixed models predicting life satisfaction from sports club participation, with twin pair IDs as random effects.

Outcome	Predictor	Sample		
		both MZ and DZ twins (n = 877 MZ pairs, 869 DZ pairs)	MZ twins only (n = 877 pairs)	MZ twins only, outcome measured at time 2 (n >= 899 individuals, see Note)
GLS	age	.03** [-.04, -.01]	-.01 [-.02, .01]	-.01 [-.01, .02]
	sex	-.09 [-.19, .02]	.07 [-.03, .18]	-.03 [-.13, .08]
	β_B	.32*** [.23, .40]	.27*** [.15, .38]	.07 [-.05, .19]
	β_W	.15** [.04, .25]	.01 [-.16, .18]	.05 [-.18, .27]
	β_{Bsat}	-	-	.75*** [.69, .82]
	β_{Wsat}	-	-	.47*** [.38, .56]
DLS: health	age	-.02*** [-.03, -.01]	-.03** [-.04, -.01]	-.00 [-.02, .02]
	sex	-.06 [-.13, .01]	-.09 [-.19, .02]	-.03 [-.15, .10]
	β_B	.12** [.04, .20]	.11 [.00, .22]	.03 [-.11, .16]
	β_W	.26*** [.14, .38]	.31*** [.13, .49]	.09 [-.16, .35]
	β_{Bsat}	-	-	.55*** [.47, .63]
	β_{Wsat}	-	-	.34*** [.25, .43]
DLS: leisure	age	-.02** [-.03, .01]	-.02* [-.17, .03]	-.00 [-.02, .02]
	sex	-.05 [-.12, .02]	-.07 [-.17, .03]	-.05 [-.18, .08]
	β_B	.20*** [.12, .29]	.23*** [.11, .34]	.14 [-.01, .28]
	β_W	.19*** [.07, .31]	.21* [.02, .40]	-.08 [-.35, .19]
	β_{Bsat}	-	-	.41*** [.33, .50]
	β_{Wsat}	-	-	.29*** [.19, .38]
DLS: friends	age	-.02** [-.03, .00]	-.01 [-.03, .00]	-.00 -.03, .02]
	sex	.01 [-.06, .07]	-.00 [-.11, .10]	-.02 [-.15, .12]
	β_B	.12** [.04, .20]	.10 [-.01, .21]	.10 [-.05, .25]
	β_W	.19** [.08, .31]	.17 [-.01, .34]	.05 [-.22, .31]
	β_{Bsat}	-	-	.46*** [.36, .55]
	β_{Wsat}	-	-	.35***

Note: β_B = between-pair effects for frequent sports club participation; β_W = within-pair effects for frequent sports club participation. For longitudinal analyses, β_{Bsat} = between-pair effects for general or domain-specific life satisfaction at time 1; β_{Wsat} = within-pair effects for satisfaction at time 1. GLS = general life satisfaction, DLS = domain-specific life satisfaction. Sample sizes for longitudinal analyses: $n = 899$ individuals for GLS, 945 for DLS: health, 932 for DLS: leisure, 927 for DLS: friends. Sample sizes differ due to missing data for some participants at the second measurement time. Intercepts are omitted for brevity.

* = $p < .05$, ** = $p < .01$, *** = $p < .001$.

Table 6

Welch's t-tests Comparing Life Satisfaction Among Twins Discordant for Frequent Sports Club Participation.

Variable	Zygoty	Twin in sports club		Discordant co-twin		df	<i>t</i>	<i>p</i>	Cohen's <i>d</i>
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>				
GLS	MZ	3.86	.85	3.86	.85	300	.08	.935	.01 [-.22, .24]
	DZ	3.91	.77	3.73	.81	524.78	2.61**	.009	.23 [.06, .40]
DLS: Health	MZ	8.30	1.53	7.58	2.39	255.55	3.09**	.002	.36 [.12, .58]
	DZ	7.95	2.13	7.47	2.24	524.7	2.65**	.008	.23 [.06, .40]
DLS: Leisure	MZ	7.85	1.92	7.36	2.32	292.4	1.93*	.0499	.23 [.00, .45]
	DZ	7.59	2.09	7.17	2.36	518.6	2.16*	.031	.19 [.02, .36]
DLS: Friends	MZ	8.60	1.76	8.27	2.13	289.92	1.47	.142	.17 [-.06, .40]
	DZ	8.61	1.6	8.18	2.18	482.27	2.59**	0.0098	.23 [.05, .40]

Note: GLS = general life satisfaction (range: 1 - 5), DLS = domain-specific life satisfaction (range: 1 - 10). For all analyses, $n = 151$ MZ twin pairs and 264 DZ pairs. Welch's *t*-tests are used because they are robust to violations of the assumption of equal variances, with little to no disadvantages compared to Student's *t*-test (Delacre et al., 2017).

5 Discussion

5.1 General discussion

The present study shows that participating in sports clubs is linked to general and domain-specific life satisfaction and that genetic and environmental confounding account for some, though not all, of these associations.

Within the utilized twin panel, we observed cross-sectional associations between adolescents' and young adults' participation in a sports club and their satisfaction with life in various domains, including general life satisfaction as well as domain-specific satisfaction with health, social life and leisure activities. We employed regression models suitable for twin samples to disentangle the effects attributable to genetic and environmental factors shared by twins from those reflecting nonshared environmental influences.

The results indicate that the relationship between sports club participation and general life satisfaction was attributable to both within-pair and between-pair effects, but this was the case only in the full sample, which included DZ twins. The within-pair effect disappeared when limiting the sample to MZ pairs, suggesting that phenotypic correlations between participation and general life satisfaction can be fully explained by genetic and environmental confounding. Comparisons within DZ pairs control for shared environmental confounders to the same extent as MZ pairs, assuming that MZ twins do not grow up in more similar environments than DZ pairs (equal environments assumption; see section 3.3.). Therefore, the presence of a significant within-pair effect in the full sample, but not among MZ twins, indicates that genetic (as opposed to shared environmental) confounding plays at least a partial role in this case.

For satisfaction with leisure, both the between-pair and within-pair effects remained significant among MZ twins and were of similar magnitude. This suggests a combination of confounding through common genetic and environmental factors as well as nonshared environmental effects beyond this confounding. Carlin et al. (2005) highlight that, in this scenario, two individuals differing in the exposure (here, sports club participation) should, on average, exhibit the same difference in outcome regardless of whether they are MZ twins or unrelated. However, for satisfaction with health, the pattern appears different, with the within-pair effect seeming to be the primary factor. Nevertheless, caution is warranted in interpreting this as evidence for nonshared environmental effects only, as the between-factor only very marginally misses statistical significance and the confidence intervals of both terms still overlap.

For satisfaction with one's social life ("networks of friends and acquaintances"), neither the within-pair effect nor the between-pair effect was statistically significant among MZ twins. Thus, it cannot be concluded that frequent participation in sports clubs is associated with this outcome beyond

genetic and shared environmental confounding. In the sample including DZ twins, these coefficients did reach significance. Interpreting this as evidence for genetic confounding, as described above, seems questionable since coefficients in both samples were almost identical numerically, and the difference in significance is clearly explained by the differing sample sizes.

None of these within-pair effects remained significant when future life satisfaction was used as the outcome, controlling for earlier levels of satisfaction. This and the remaining possibility of nonshared environmental confounders mean the present results should not be taken as conclusive evidence of causal links (see Limitations).

These patterns of results are reflected in descriptive differences within pairs discordant for sports club participation. Welch's t-tests showed that among 151 MZ twins pairs where only one twin regularly participated in a sports club, that twin was, on average, significantly happier about their health and their leisure time activities than their co-twin. Effect sizes were small (Cohen's $d = 0.36$ and 0.23 , respectively; see Cohen, 1988).

Previous research has established physical activity in general as a possible protective factor for mental and physical health and as a precursor to life satisfaction. Some studies have indicated particularly pronounced effects of sports clubs as opposed to other forms of physical activity, perhaps because of their social aspect. The present results add to this literature by providing some tentative evidence that these associations remain solid, although small in effect size, when genetic and shared environmental confounding is controlled for. To our knowledge, it is the first study examining within-pair effects among twins in relation to sports club participation. Two previous studies applied similar analyses to measures of physical activity and life satisfaction: Waller et al. (2010) found significant within-pair effects, providing evidence of associations beyond genetic and shared environmental confounding, in a longitudinal study that predicted general life satisfaction from physical activity. Stubbe et al. (2007) found no significant differences in general life satisfaction among MZ twins discordant for regular exercise. The different scope of these studies and the present one make it difficult to arrive at clear-cut conclusions, and further research is necessary to establish if, and in what context, physical activity directly benefits general or domain specific life satisfaction. If sports clubs do have a causal effect on well-being, is this effect driven by the physical activity itself, or does the social aspect have additional merit, as suggested by some studies in the general population (e. g., Vilhjalmsson & Thorlindsson, 1992)? The fact that associations to domain-specific satisfaction with one's physical health were the most pronounced in our study suggests that staying physically fit could play a role, but this is not necessarily purely an effect of physical activity since social connectedness may also improve physical and mental health (Haslam et al., 2015).

A suitable comparison category would be other clubs for leisure activities, unrelated to sports, where adolescents could be active. Participants in TwinLife, the family panel dataset used in the present study, were indeed asked not only about sports clubs, but various kinds of leisure associations, but these included a highly heterogeneous range of activities, from religious communities to becoming a voluntary firefighter. Sports clubs were the most frequent activity in the sample by a wide margin, while the other activities were mostly scattered across relatively few twin pairs. For these reasons, a direct comparison did not seem viable within the present study.

In addition to the main analysis, a secondary finding of the present study was the substantial stability in environmentally mediated sources of variance in all facets of life satisfaction: within-pair coefficients were almost as important in predicting future satisfaction among MZ twins as between-pair effects. Differences between MZ twins in life satisfaction therefore appear to be meaningful and partially stable, encouraging future investigations into the specific nonshared environmental factors at play. Generally speaking, variables of well-being, such as life satisfaction, are among the traits with the most pronounced “E” components in univariate twin models and are thus promising candidate variables in the search for nonshared environmental influences on psychological traits.

5.2 Limitations

There are several limitations to our study that we would like to acknowledge. First of all, the cross-sectional links we found are far from establishing causation. If such a causal link exists, it might work in the opposite direction: happier adolescents might be more inclined to join a sports club. The concurrent nature of our main analyses cannot rule this out. In tentative longitudinal analyses, within-pair associations did not show any incremental validity in predicting subsequent life satisfaction beyond earlier satisfaction. This is consistent with life satisfaction as a possible cause of participation rather than the other way around. It must be noted that our longitudinal results were based on smaller sample sizes, particularly within the subset of discordant twins. With one exception, the nonsignificant coefficients for sports club participation were numerically positive, hinting at possible small effects that remain, but would require far more statistical power to investigate conclusively.

Furthermore, while comparing MZ twins implicitly rules out genetic and shared environmental explanations, nonshared environmental influences can still act as confounding variables and lead to spurious within-pair effects (Frisell et al., 2012). If a third variable is the true cause for the observed exposure, such as sports club participation in the present study, discordant twin pairs will typically be discordant on the confounding variable as well, which biases within-pair comparisons. For example, if one of the twins in an MZ pair is physically healthier, they might both be more likely to

join a sports club and to be more satisfied with life. Since the present analysis can only rule out common genetic and environmental background, such a difference between MZ twins in a third variable would not be controlled for.

Due to the panel nature of the present study, very brief measures were utilized, including single items in the case of domain-specific life satisfaction. Focusing on brevity in panel studies allows the inclusion of a wider range of measures, but comes with the tradeoff of lower reliability and, in turn, validity. This is particularly true as twin difference models are prone to bias due to measurement error (Frisell et al., 2012). This low scale quality is reflected in the fact that parameter estimates of variance components for the measures of life satisfaction showed nonshared environmental contributions (which include measurement error) of up to .84. This was particularly pronounced for the single-item measures, while the 5-item scale for general life satisfaction, which showed very good internal consistency, generated lower estimates of E. Previous studies on the topic have also shown substantial nonshared environmental components to life satisfaction (Bartels et al., 2015), but these coefficients are still unusually large and presumably reflect relatively large measurement error. Notably, the best-fitting models for the domain-specific items on satisfaction with health and friends suggested non-additive genetic effects (also found by Hahn et al., 2013).

Our measurement of participation is very broad. Sports clubs may differ in numerous aspects, such as the particular sport being practiced. A comparison of characteristics of sports club across European countries reveals marked differences, such as German clubs placing relatively low emphasis on competitiveness and success (Breuer et al., 2015b). This makes it unclear to what extent our results can be generalized beyond the current sample or the country of Germany. It also stands to reason that the adolescents and young adults in the sample may have highly heterogeneous motives for participating, and that some of them are urged to do so by their parents rather than joining a sports club out of intrinsic motivation. Our analysis cannot differentiate between such contextual differences. The fact that concordance within twin pairs declined with age appears to be consistent with the common sense assumption that joining a club becomes more of an individual decision in adolescence and young adulthood, as parents have less and less influence over their offspring's life choices. It might thus start to more strongly reflect a person's heritable tendencies (e.g. through active gene-environment correlation), making it less likely that sports participation impacts life satisfaction through an environmentally mediated pathway. This remains to be explored by future genetically informed research. It also remains unclear if there is a dose-response relationship between sports club participation and well-being, i.e., whether more frequent participation is more beneficial than occasional activity (as is the case for physical activity in general, e.g. Lee & Skerrett, 2001). In the present study, self-reported participation was used as a

dichotomized variable to ease interpretation and because most twins either reported frequent participation or none at all. Studies that use more refined scales to measure the extent and nature of participation in sports club could yield more detailed conclusions.

Despite these limitations, the findings which the present study presents based on a large, representative twin sample suggest that sporting activities with peers may play a role in impacting life satisfaction. Specific factors that are associated with life satisfaction, or other measures of well-being, through nonshared environmental pathways have remained difficult to uncover in behavioral genetic research to date (see Turkheimer & Waldron, 2000; Plomin et al., 2001; Smith, 2011).

Given this context, the finding that significant effects of sports clubs participation on well-being persist when genetic and shared environmental confounding is controlled for is notable. It suggests that adolescents and young adults who participate in sports clubs may, on average, be happier not only because they are, for example, genetically predisposed to high life satisfaction regardless. While the present research cannot conclusively demonstrate that joining a sports club causally impacts happiness among adolescents and young adults, it encourages offering such opportunities to them.

References

- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317-332.
- Arbet, J., McGue, M., & Basu, S. (2020). A robust and unified framework for estimating heritability in twin studies using generalized estimating equations. *Statistics in medicine*, 39(27), 3897-3913. <https://doi.org/10.1002/sim.8564>
- Arseneault, L., Milne, B. J., Taylor, A., Adams, F., Delgado, K., Caspi, A., & Moffitt, T. E. (2008). Being bullied as an environmentally mediated contributing factor to children's internalizing problems: a study of twins discordant for victimization. *Archives of pediatrics & adolescent medicine*, 162(2), 145-15.
- Bartels, M. (2015). Genetics of wellbeing and its components satisfaction with life, happiness, and quality of life: A review and meta-analysis of heritability studies. *Behavior genetics*, 45, 137-156.
- Bates, T. C., Maes, H., & Neale, M. C. (2019). umx: twin and path-based structural equation modeling in R. *Twin Research and Human Genetics*, 22(1), 27-41.
- Becchetti, L., Giachin Ricca, E., & Pelloni, A. (2012). The relationship between social leisure and life satisfaction: Causality and policy implications. *Social indicators research*, 108, 453-49.
- Breuer, C., Feiler, S., & Wicker, P. (2015a). Sport clubs in Germany. *Sport clubs in Europe: A cross-national comparative perspective*, 187-208.
- Breuer, C., Hoekman, R., Nagel, S., & van der Werff, H. (2015b). *Sport clubs in Europe* (Vol. 12). Heidelberg, Germany: Springer.
- Buecker, S., Simacek, T., Ingwersen, B., Terwiel, S., & Simonsmeier, B. A. (2021). Physical activity and subjective well-being in healthy individuals: a meta-analytic review. *Health Psychology Review*, 15(4), 574-592.
- Carlin, J. B., Gurrin, L. C., Sterne, J. A., Morley, R., & Dwyer, T. (2005). Regression models for twin studies: a critical review. *International journal of epidemiology*, 34(5), 1089-1099.
- Carlsson, S., Andersson, T., Lichtenstein, P., Michaelsson, K., & Ahlbom, A. (2006). Genetic effects on physical activity: results from the Swedish Twin Registry. *Medicine and Science in Sports and Exercise*, 38(8), 1396-1401.
- Cerin, E., Leslie, E., Sugiyama, T., & Owen, N. (2009). Associations of multiple physical activity domains with mental well-being. *Mental Health and Physical Activity*, 2(2), 55-64.
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. 2nd edn. Hillsdale, NJ: Lawrence Erlbaum.
- Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology*, 30(1), 92-101.

- Diener, E., Lucas, R. E., & Oishi, S. (2018). Advances and open questions in the science of subjective well-being. *Collabra: Psychology*, 4(1).
- Diener, E., & Ryan, K. (2009). Subjective well-being: A general overview. *South African journal of psychology*, 39(4), 391-406.
- Duncan, G. E., Goldberg, J., Noonan, C., Moudon, A. V., Hurvitz, P., & Buchwald, D. (2008). Unique environmental effects on physical activity participation: a twin study. *PLoS One*, 3(4), e2019.
- Eime, R. M., Harvey, J. T., Brown, W. J., & Payne, W. R. (2010). Does sports club participation contribute to health-related quality of life. *Med Sci Sports Exerc*, 42(5), 1022-8.
- Elmose-Østerlund, K., Seippel, Ø., Llopis-Goig, R., Van der Roest, J. W., Adler Zwahlen, J., & Nagel, S. (2019). Social integration in sports clubs: Individual and organisational factors in a European context. *European Journal for Sport and Society*, 16(3), 268-29.
- Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics*, 29(1), 51-76.
- Felson, J. (2014). What can we learn from twin studies? A comprehensive evaluation of the equal environments assumption. *Social science research*, 43, 184-199.
- Fisher, A., van Jaarsveld, C. H., Llewellyn, C. H., & Wardle, J. (2010). Environmental influences on children's physical activity: quantitative estimates using a twin design. *PloS One*, 5(4), e1011.
- Fox, K. R. (1999). The influence of physical activity on mental well-being. *Public health nutrition*, 2(3a), 411-418.
- Franzen, A., & Botzen, K. (2011). Vereine in Deutschland und ihr Beitrag zum Wohlstand der Regionen. [Associations in Germany and their contribution to local prosperity.] *Soziale Welt*, 391-413.
- Frisell, T., Öberg, S., Kuja-Halkola, R., & Sjölander, A. (2012). Sibling comparison designs: Bios from non-shared confounders and measurement error. *Epidemiology*, 713-72.
- Gidziela, A., Malanchini, M., Rimfeld, K., McMillan, A., Ronald, A., Viding, E., ... & Plomin, R. (2023). Explaining the influence of non-shared environment (NSE) on symptoms of behaviour problems from preschool to adulthood: mind the missing NSE gap. *Journal of child psychology and psychiatry*, 64(5), 747-757.
- Hahn, E., Johnson, W., & Spinath, F. M. (2013). Beyond the heritability of life satisfaction—The roles of personality and twin-specific influences. *Journal of Research in Personality*, 47(6), 757-767.
- Haslam, C., Cruwys, T., Haslam, S. A., & Jetten, J. (2015). Social connectedness and health.

Encyclopaedia of geropsychology, 2015, 46-1.

- Irwin, J. R., & McClelland, G. H. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, 40(3), 366-371.
- Jose, P. E., Ryan, N., & Pryor, J. (2012). Does social connectedness promote a greater sense of well-being in adolescence over time? *Journal of research on adolescence*, 22(2), 235-251.
- Korhonen, T., Kujala, U. M., Rose, R. J., & Kaprio, J. (2009). Physical activity in adolescence as a predictor of alcohol and illicit drug use in early adulthood: a longitudinal population-based twin study. *Twin research and human genetics*, 12(3), 261-268.
- Kujala, U. M., Kaprio, J., Sarna, S., & Koskenvuo, M. (1998). Relationship of leisure-time physical activity and mortality: the Finnish twin cohort. *Jama*, 279(6), 440-444.
- Lang, V., & Kottwitz, A. (2020). The socio-demographic structure of the first wave of the TwinLife panel study: A comparison with the microcensus. *methods, data, analyses*, 14(1), 28.
- Larsen, M. N., Madsen, M., Cyril, R., Madsen, E. E., Lind, R. R., Ryom, K., ... & Krstrup, P. (2021). Well-being, physical fitness and health profile of 10–12 years old boys in relation to leisure-time sports club activities: A cross-sectional study. *BMJ open*, 11(11), e050194.
- Lent, R. W., Singley, D., Sheu, H. B., Gainor, K. A., Brenner, B. R., Treistman, D., & Ades, L. (2005). Social cognitive predictors of domain and life satisfaction: Exploring the theoretical precursors of subjective well-being. *Journal of counseling psychology*, 52(3), 429.
- Leskinen, T., Sipilä, S., Alen, M., Cheng, S., Pietiläinen, K. H., Usenius, J. P., ... & Kujala, U. M. (2009). Leisure-time physical activity and high-risk fat: a longitudinal population-based twin study. *International Journal of Obesity*, 33(11), 1211-1218.
- Long, R. F., & Huebner, E. S. (2014). Differential validity of global and domain-specific measures of life satisfaction in the context of schooling. *Child Indicators Research*, 7, 671-694.
- Lüdecke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, 5(53), 2445.
- Maher, B. (2008). The case of the missing heritability. *Nature*, 456(7218), 18-22.
- Maher, J. P., & Conroy, D. E. (2017). Daily life satisfaction in older adults as a function of (in) activity. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 72(4), 593-602.
- McAdams, T. A., Rijdsdijk, F. V., Zavos, H. M., & Pingault, J. B. (2021). Twins and causal inference: leveraging nature's experiment. *Cold Spring Harbor perspectives in medicine*, 11(6), a039552.
- McGue, M., & Bouchard Jr, T. J. (1984). Adjustment of twin data for the effects of age and sex.

Behavior genetics, 14(4), 325-343.

- McGue, M., Osler, M., & Christensen, K. (2010). Causal inference and observational research: The utility of twins. *Perspectives on psychological science*, 5(5), 546-556.
- Misheva, V. (2016). What determines emotional well-being? The role of adverse experiences: Evidence using twin data. *Journal of happiness studies*, 17, 1921-1937.
- Mönkediek, B., Lang, V., Weigel, L., Baum, M. A., Eifler, E. F., Hahn, E., ... & Spinath, F. M. (2019). The German twin family panel (TwinLife). *Twin Research and Human Genetics*, 22(6), 540-547.
- Mustelin, L., Joutsu, J., Latvala, A., Pietiläinen, K. H., Rissanen, A., & Kaprio, J. (2012). Genetic influences on physical activity in young adults: a twin study. *Medicine and science in sports and exercise*, 44(7), 1293-1301.
- Mustelin, L., Silventoinen, K., Pietiläinen, K., Rissanen, A., & Kaprio, J. (2009). Physical activity reduces the influence of genetic effects on BMI and waist circumference: a study in young adult twins. *International journal of obesity*, 33(1), 29-36.
- Mutz, M., Reimers, A. K., & Demetriou, Y. (2021). Leisure time sports activities and life satisfaction: deeper insights based on a representative survey from Germany. *Applied Research in Quality of Life*, 16, 2155-2171.
- Nes RB, Czajkowski N, Roysamb E, Reichborn-Kjennerud T, Tambs K (2008). Well-being and ill-being: shared environments, shared genes? *Journal of Positive Psychology*, 3(4):253–265
- O'Rourke, H. M., & Sidani, S. (2017). Definition, determinants, and outcomes of social connectedness for older adults: a scoping review. *Journal of Gerontological Nursing*, 43(7), 43-52.
- Pedišić, Ž., Greblo, Z., Phongsavan, P., Milton, K., & Bauman, A. E. (2015). Are total, intensity- and domain-specific physical activity levels associated with life satisfaction among university students? *PloS one*, 10(2), e0118137.
- Pietsch, S., Linder, S., & Jansen, P. (2022). Well-being and its relationship with sports and physical activity of students during the coronavirus pandemic. *German Journal of Exercise and Sport Research*, 52(1), 5.
- Pike, A., Reiss, D., Hetherington, E. M., & Plomin, R. (1996). Using MZ differences in the search for nonshared environmental effects. *Journal of Child Psychology and Psychiatry*, 37(6), 695-704.
- Pingault, J. B., O'reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijdsdijk, F., & Dudbridge, F. (2018). Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics*, 19(9), 566-58.

- Plomin, R. (2011). Commentary: Why are children in the same family so different? nonshared environment three decades later. *International journal of epidemiology*, 40(3), 582-592.
- Plomin, R., Asbury, K., & Dunn, J. (2001). Why are children in the same family so different? Nonshared environment a decade later. *The Canadian Journal of Psychiatry*, 46(3), 225-233.
- Plomin, R., & Daniels, D. (1987). Why are children in the same family so different from one another? *Behavioral and brain Sciences*, 10(1), 1-16.
- Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2016). Top 10 replicated findings from behavioral genetics. *Perspectives on psychological science*, 11(1), 3-23.
- Polderman, T. J., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics*, 47(7), 702-709.
- Rijsdijk, F. V., & Sham, P. C. (2002). Analytic approaches to twin data using structural equation models. *Briefings in bioinformatics*, 3(2), 119-133.
- R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Rowe, D. C., & Plomin, R. (1981). The importance of nonshared (E²) environmental influences in behavioral development. *Developmental Psychology*, 17(5), 517.
- Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegate, H., Teplitsky, C., ... & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in ecology and evolution*, 11(9), 1141-1152.
- Schmidt, S. C., Anedda, B., Burchartz, A., Oriwol, D., Kolb, S., Wäsche, H., ... & Woll, A. (2020). The physical activity of children and adolescents in Germany 2003-2017: The MoMo-study. *PLoS one*, 15(7), e0236117.
- Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of clinical epidemiology*, 98, 146-151.
- Schmiedeborg, C., & Schröder, J. (2017). Leisure activities and life satisfaction: An analysis with German panel data. *Applied Research in Quality of Life*, 12, 137-151.
- Schnittker, J. (2008). Happiness and success: Genes, families, and the psychological effects of socioeconomic position and social support. *American Journal of Sociology*, 114(S1), S233-S259.
- Sjölander, A., Frisell, T., & Öberg, S. (2012). Causal interpretation of between-within models for twin research. *Epidemiologic Methods*, 1(1), 217-237.
- Smith, G. D. (2011). Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. *International Journal of Epidemiology*, 40(3), 537-

- Spengler, S., & Woll, A. (2013). The more physically active, the healthier? The relationship between physical activity and health-related quality of life in adolescents: the MoMo study. *Journal of physical activity and health*, 10(5), 708-715.
- Steger, M. F., & Kashdan, T. B. (2007). Stability and specificity of meaning in life and life satisfaction over one year. *Journal of Happiness Studies*, 8, 161-179.
- Streiner, D. L. (2002). Breaking up is hard to do: the heartbreak of dichotomizing continuous data. *The Canadian Journal of Psychiatry*, 47(3), 262-266.
- Stubbe, J. H., de Moor, M. H., Boomsma, D. I., & de Geus, E. J. (2007). The association between exercise participation and well-being: a co-twin study. *Preventive medicine*, 44(2), 148-152.
- Tikhodeyev, O. N., & Shcherbakova, O. V. (2019). The problem of nonshared environment in behavioral genetics. *Behavior genetics*, 49(3), 259-269.
- Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current directions in psychological science*, 9(5), 160-164.
- Turkheimer, E., & Waldron, M. (2000). Nonshared environment: a theoretical, methodological, and quantitative review. *Psychological bulletin*, 126(1), 78.
- Varca, P. E., Shaffer, G. S., & Saunders, V. (1984). A Longitudinal Investigation of Sport Participation and Life Satisfaction. *Journal of Sport Psychology*, 6(4).
- Veenhoven, R. (1996). The study of life satisfaction. In W. E. Saris, R. Veenhoven, A. C. Scherpenzeel, & B. Bunting (Eds.), *A comparative study of satisfaction with life in Europe* (pp. 1–37). Budapest: Eötvös University Press.
- Veliz, P., Snyder, M., & Sabo, D. (2019). *The State of High School Sports in America: An Evaluation of the Nation's Most Popular Extracurricular Activity*. New York, NY: Women's Sports Foundation.
- Vilhjalmsson, R., & Thorlindsson, T. (1992). The integrative and physiological effects of sport participation: A study of adolescents. *The Sociological Quarterly*, 33(4), 637-647.
- Vink, J. M., Boomsma, D. I., Medland, S. E., de Moor, M. H., Stubbe, J. H., Cornes, B. K., ... & de Geus, E. J. (2011). Variance components models for physical activity with age as modifier: a comparative twin study in seven countries. *Twin Research and Human Genetics*, 14(1), 25-34.
- Vitaro, F., Brendgen, M., & Arseneault, L. (2009). The discordant MZ-twin method: One step closer to the holy grail of causality. *International Journal of Behavioral Development*, 33(4), 376-382.
- Waller, K., Kujala, U., Kaprio, J., Koskenvuo, M., & Rantanen, T. (2010). Effect of physical activity on health in twins: A 30-yr longitudinal study. *Medicine & Science in Sports &*

Exercise, 42(4).

Wicker, P., & Frick, B. (2015). The relationship between intensity and duration of physical activity and subjective well-being. *The European Journal of Public Health*, 25(5), 868-872.

Wiese, C. W., Kuykendall, L., & Tay, L. (2018). Get active? A meta-analysis of leisure-time physical activity and subjective well-being. *The Journal of Positive Psychology*, 13(1), 57-66.

Supplementary Materials

Table S1. *Sample characteristics by birth cohort*

Variables	Cohort 3				Cohort 4			
	MZ		DZ		MZ		DZ	
	m	f	m	f	m	f	m	f
n	376	470	408	558	364	544	334	438
age	17.04	17.00	17.00	17.03	23.05	23.04	22.98	23.03
sports club participation ^a	(.37)	(.39)	(.32)	(.31)	(.86)	(.81)	(.75)	(.81)
	.61	.50	.57	.52	.48	.37	.42	.29
GLS (time 1)	3.98	3.89	3.91	3.79	3.79	3.95	3.71	3.81
	(.78)	(.78)	(.73)	(.85)	(.77)	(.76)	(.79)	(.81)
DLS: health (time 1)	8.32	7.93	7.99	7.77	7.81	7.74	7.51	7.52
	(1.90)	(2.38)	(2.11)	(2.30)	(1.91)	(2.15)	(2.20)	(2.30)
DLS: leisure (time 1)	8.06	7.48	7.65	7.55	7.37	7.51	7.31	7.22
	(1.95)	(2.27)	(2.27)	(2.11)	(2.24)	(2.18)	(2.08)	(2.26)
DLS: friends (time 1)	8.29	8.31	8.34	8.27	8.35	8.17	7.90	8.27
	(1.87)	(1.80)	(1.76)	(1.67)	(1.47)	(1.99)	(1.76)	(1.78)
GLS (time 2)	3.93	3.83	3.83	3.81	3.83	3.91	3.74	3.81
	(.79)	(.79)	(.76)	(.75)	(.77)	(.84)	(.80)	(.81)
DLS: health (time 2)	7.84	7.55	7.70	7.49	(7.61)	7.52	7.43	7.26
	(2.24)	(2.47)	(2.03)	(2.34)	1.87	(2.14)	(2.12)	(2.47)
DLS: leisure (time 2)	7.50	7.23	7.67	7.24	7.30	7.16	7.05	7.09
	(2.07)	(2.20)	(2.03)	(2.11)	(1.90)	(2.21)	(2.26)	(1.94)
DLS: friends (time 2)	8.72	8.48	8.54	8.45	8.31	8.47	8.15	8.27
	(1.64)	(1.88)	(1.83)	(1.95)	(1.82)	(1.88)	(1.91)	(1.78)

Note: ^a 0 = no frequent participation (includes “never”, “less than once a month” and “monthly”; see section 3.2) and 1 = weekly participation. GLS = general life satisfaction (range: 1 - 5), DLS = domain-specific life satisfaction (range: 1 - 10). Sample sizes refer to the number of individuals, not of twin pairs. Standard deviations are given in brackets.

Cohorts 1 and 2 are not included in the analyses and are omitted for brevity.

Beyond case vignettes: Do diagnostic labels affect how symptoms of mental illness are perceived?

Dings, Alexander*^a, Kubillus, Lara K.^a

^aUniversität des Saarlandes, Saarbrücken, Germany

*Corresponding author; e-mail address: alexander.dings@uni-saarland.de

Abstract

A long-standing question in clinical psychology concerns the merits and drawbacks of using diagnostic labels in communication about mental illness. A frequent view is that such labels contribute to mental illness stigma, for example by fostering a categorical view of affected people as part of a distinct out-group. Previous research has found some evidence to support such claims, but results are mixed. The current study investigates the effects of diagnostic labels at the level of specific behaviors, rather than subsuming them in a case vignette, in a sample of psychology undergraduates in Germany. Results indicate that when behaviors are presented along with a diagnostic label, participants perceive them as less common, less blameworthy and less typical for themselves, but not as less acceptable, normal, treatable or dependent on outside help. These results held true for behaviors across both actual DSM-5 symptoms of mental illnesses and similar behaviors that were not genuine diagnostic criteria. Labeling effects were generally similar across heterogeneous diagnoses and items. Factor analyses showed that the different behaviors associated with a particular diagnosis were rated more uniformly when they were accompanied by a diagnostic label than when they were not.

Introduction

Being afflicted with a mental illness remains a highly stigmatized condition across different cultures [1, 2]. In a highly influential overview on the topic published in 2005, Rüsch and colleagues [3] noted, among other things, what mental illness stigma typically encompasses: frequent prejudices include overestimating the prevalence of violent behavior among people with a mental illness [4] or the belief that those with a mental illness will usually not be able to overcome it [5]. As a consequence, most people with a mental illness report experiencing some form of discrimination [6,7].

Accordingly, they are often cautious about when and to whom to disclose their diagnosis [8], especially because such a disclosure cannot easily be backtracked fully [9]. Such disclosure can result in more experiences of discrimination [10], but evidence is mixed and suggests that disclosure can also have positive consequences such as increased social support [11, 12].

A prominent hypothesis in the literature on mental illness stigma proposes that part of the problem in this regard is the practice of labeling mental illness with diagnostic terms. Conceptualizing people with a mental illness as an out-group is seen as a central aspect of stigmatization [3]. Categorical diagnostic labels could facilitate such thinking. There is some evidence that thinking of mental illness as a continuum rather than a categorical distinction is associated with less stigma [13], supporting the idea that labels might do more harm than good as they may run counter to such a continuum model.

This argument touches on the more general idea known as the Sapir-Whorf hypothesis, first formulated in 1929 [14], according to which language shapes human perception rather than merely being a vehicle for communicating said perception. Under such a framework, it is logical to assume that the language used to talk about marginalized groups matters because it helps to enforce or alleviate the stereotypes that form the basis for stigmatization.

In this vein, Rüsch et al. [3] argued that “language can be a powerful source and sign of stigmatization” (p. 530), pointing out a stark difference between describing physical and mental ailments: it is common to describe a person with schizophrenia as “schizophrenic”, but the equivalent of calling a cancer patient “cancerous” would be highly unusual and inappropriate. Most self-help groups and other advocates for people with mental illnesses accordingly recommend so-called person-first language (such as “a person with schizophrenia”). However, evidence is mixed on whether using person-first language actually reduces stigma, with some studies finding no effect [15], and there are also some authors who see it as misguided [16]. A large-scale study in 2021 [17] found that using different terms for mental illness – such as the strong word “disorder” - had unexpectedly small consequences on participants' attitudes. These results cast doubt on the general

importance of word choice. Given these results, labeling effects should not be assumed to be true in the absence of solid empirical evidence, and currently, the effects of using diagnostic labels are still not entirely clear.

The idea that labeling mental illnesses with diagnostic terms aggravates stigmatization was most notably introduced in 1966 by Scheff [18], who went so far as to suggest that diagnostic labels are a central reason why symptoms of mental illnesses are even perceived negatively as disorders rather than fully acceptable deviations from a social norm. Over time, a more moderate view became the consensus: although Scheff's approach was deemed too radical, studies in the following decades did nevertheless show that labeling effects are likely real and impactful [19]. As a more recent illustrative example, Altmann and colleagues [20] showed that using a label when describing relatively mild, marginal cases of mental illness increases perceptions of those symptoms as more stable and less controllable compared to when no label is used. These developments were synthesized into a "modified labeling approach" by Link and colleagues in 1987 [21], which states that although "labeling does not directly produce mental disorder" (p. 400), it nevertheless shapes how especially people with a mental illness view themselves, i.e., labels can lead to self-stigma. The assumption that labels aggravate stigma, either towards oneself or others, has now been part of the mental illness stigma literature for decades and "has almost reached the status of a cultural truism" [22 p 461]. At the same time, empirical evidence is still inconclusive on the extent and nature of labeling effects. In the only meta-analysis to date, O'Connor and colleagues [23] reviewed the existing literature and identified 22 studies matching their inclusion criteria. Results differed notably by diagnosis, leading the authors to conclude that "labeling effects cannot meaningfully be discussed in general 'catch-all' terms" [p 7]. Two different overviews are worth mentioning: Sims and colleagues [24] summarized the various psychosocial consequences of receiving a medical diagnosis, including a psychiatric label. However, this review focused on qualitative studies, not on experimental tests of differences between the presence or absence of a label. Despite this, it is noteworthy that results under this qualitative framework are mixed as well. In addition, Franz and colleagues [25] reviewed experimental studies focused specifically on the school context and found that mentioning a diagnosis such as, for example, ADHD was often associated with worse evaluations by teachers. There is therefore better evidence for labeling effects in this particular context than for mental illness labels in general. However, this meta-analysis has similar issues as mentioned before: most studies were vignette-based, and results varied considerably by diagnosis.

One reason for these diverse findings is the heterogeneity in approaches. There are countless possible moderators in mental illness stigma research in general – results may differ depending on,

among others, which diagnoses, respondents, outcome variables, or culture researchers focus on. Besides these general issues, labeling effects have been investigated using different study designs, with different strengths and drawbacks. For example, studying inferred labels needs to be distinguished from causal effects of assigned labels. A 2003 study [26] that became a frequently cited source of the importance of labeling effects presented participants with unlabeled case vignettes and compared their perception of the person depending on which diagnosis, if any, the participants ascribed to them. While informative, this approach, as well as more recent similar studies [27] does not represent experimental evidence for labeling effects, as participants are not randomly allocated to labeling conditions but introduce the labels themselves (“inferred labels”), possibly leading to confounding effects with previous knowledge about mental illness.

While the literature on labeling effects is generally diverse, most studies do not differ in employing case vignettes to study them. We argue that there is a need for additional and different approaches, in particular because the ecological validity of case vignettes is unclear. The current study was therefore conducted with the aim of testing labeling effects in a more specific context.

A vignette necessarily presents a general overview of a person's symptoms, presenting them in a summarized way that might not reflect the way one gets to know a person in real life. In their review, O'Connor and colleagues [23, p 8] note that using them “comes at the cost of ecological validity”. In everyday interactions, contact with the symptoms of mental illness might often happen one at a time. Meeting a person with ADHD, for example, might entail being confronted with one particular behavior that stands out negatively, i. e., being unpunctual for a meeting. A labeling effect might then change one's reaction to this particular behavior if one knows, or does not know, of this person's diagnosis. The label may amplify an attitude of not wanting to interact with “this kind” of person, but it also appears plausible that it might induce more benevolent reactions by making the behavior understandable. Case vignettes operate on a broader level, inducing a more general picture of a person. They also typically introduce their subject as a person who has mental health issues, is visiting a psychiatrist etc. Even if a vignette does not make this explicit, it is typically obvious to the reader that the vignette is describing a person with a general mental health problem. On the other hand, encountering a single, specific symptom of mental illness – such as unpunctuality in ADHD – might not evoke such associations at all, with the consequence that the behavior is judged under a different framing than when it is part of a case vignette.

Martinez et al. [28] found that introducing a person as having a mental illness induced greater social distance (i.e., higher stigma) compared to mentioning a physical ailment – but the opposite effect occurred when the information was embedded into a story of a person engaging in a number of ambiguous, but somewhat rude behaviors towards others. In other words, as standalone information,

a mental illness label provoked negative associations; but in the context of explaining why a person behaves a certain way, the label appeared to be helpful as it put a person's behavior into perspective and made it more understandable, and less blameworthy.

We are not suggesting by any means that case vignettes are not a valuable and informative way to study labeling effects, but propose that it is worthwhile to add to those findings an approach that investigates stigmatizing effects at a more concrete level of behavior, especially given the heterogeneity of previous findings as pointed out by O'Connor et al. [23].

Imhoff [22] similarly highlighted the need for studies that focus on comparing perceptions of concrete behavior while experimentally manipulating the presence or absence of a diagnostic label. Using such an approach to test labeling effects on the perception of symptoms of schizophrenia in a large sample, Imhoff [22] then showed that the presence of a label was indeed associated with slightly more negative perceptions of a person in a case vignette that focused on concrete symptoms. The current study aims to conceptually replicate and further develop this approach by presenting symptoms individually instead of embedding them into the wider context of a case vignette.

The main research question of the current study is whether labeling effects might be specific to introducing individuals with mental illness in a vignette-based format or if they persist in an approach that present specific symptoms individually.

Methods and materials

Participants were recruited among psychology students at Saarland University, Germany. Besides restricting participation to that group, there were no exclusion criteria for participation. The students received course credit for their participation. All materials were presented online and in German. The data reported here were collected as part of a more extensive survey on attitudes towards mental health and mental illness. This survey was introduced as dealing with “the topic of mental health” in both groups, with no mentioning of labeling effects or stigma. The current study constituted the first part of that larger survey, i. e., no other materials were shown before the materials presented here to avoid disclosing to the participants that the study investigated stigmatizing attitudes.

Participants were randomly allocated to the “label” and “no label” experimental conditions. Instead of describing clinical symptoms in a case vignette, they were presented as standalone items with no explicit connection to each other and no introduction of a specific person. As detailed in the Methods section, this represented a novel approach to measuring labeling effects, introduced with

the goal of investigating whether such effects persist even when specific behaviors rather than more extensive case vignettes are the focus of external judgments.

Each item described a behavior corresponding to a symptom used in the DSM-5 (American Psychological Association, 2013)¹⁴ [29] for diagnosing or describing a certain mental illness. Descriptions in the DSM were the basis for phrasing the items, but were not used verbatim to reduce the risk of participants recognizing them as diagnostic criteria even in the no label condition. Instead, the phrasing of the items used colloquial language with as few specific terms from clinical psychology as possible.

Both authors phrased candidate items and narrowed them down in a joint process which was guided by several criteria. Symptoms should ideally be specific to one particular diagnosis to ensure that the associations it evoked in the no label condition were similarly narrow to those in the label condition. In other words, we aimed to avoid a situation where participants in the no label condition could react to a symptom in multiple different ways depending on which particular diagnosis it evoked for them. This was inherently not perfectly feasible: even if a symptom is an official criterion only for one particular diagnosis, associations with other disorders are inevitable due to frequent comorbidity. In the DSM-5, 37% of symptoms appear more than once, including 85% of symptoms associated with depressive disorders [30].

Another goal in item construction was to select behaviors that we assumed most people to be familiar with in a milder form. As we hypothesized labeling effects to be particularly relevant in those cases where an afflicted person's symptoms are not strikingly different from “normal” behavior, but rather represent the extreme ends of a continuum, we focused on symptoms suitable to that notion. For example, in anorexia nervosa, focusing on one's body weight and trying to reduce it was deemed a more fitting symptom for our study than, for example, purging behaviors. Such a more extreme symptom is less nuanced, presumably less relatable for a person not afflicted with the disorder (in addition to occurring only in some cases of anorexia nervosa), and is more likely to evoke an association like “eating disorder” in any case, meaning that there might not be a true “no label” condition for such “extreme” symptoms. This also helped making these items as similar as possible to the pseudo-items (see below). Lastly, we aimed to avoid symptoms that can manifest themselves in opposite directions in the same disorder (i. e., reduced or heightened need for sleep in depression, very low or very high reactivity to sensory input in autism spectrum disorders). Items in the final survey included symptoms of major depression (6 symptoms), social anxiety (4), schizophrenia (3), anorexia nervosa (3), and autism spectrum disorders (3). These are well-known

¹⁴ Note that data collection in this study took place in 2022, before the publication of the latest text revision of the DSM-5, which was published in September 2022 of that year.

diagnoses that cover a range from diagnoses with relatively low to relatively high levels of stigma [31]. For example, one of the items pertaining to autism spectrum disorders was “Someone feels uncomfortable with changes in their everyday life that require them to change their usual routine”, reflecting in more colloquial language the diagnostic criterion B2 for autism spectrum disorders in the DSM [29]. The different number of items per diagnosis reflects the fact that creating non-redundant items that fulfilled the aforementioned construction principles was more feasible for some diagnoses than for others.

In addition to the items that reflected actual DSM-5 symptoms, we also added a pseudo-item for each diagnosis (two for major depression). These were constructed with the aim of describing behaviors that resemble actual symptoms of the relevant disorder, but are not listed among the official criteria, presumably because they are too common and universal to be selective. For example, the pseudo-item for autism spectrum disorders was “Someone constantly interrupts others and does not seem to listen properly”, as this behavior is close to the difficulties in social reciprocity relevant for diagnosing autism spectrum disorders, but is frequently seen in people without such a disorder as well and does not constitute an explicit diagnostic criterion in the DSM-5. If a diagnostic labeling effect exists and leads to “normal” behavior acquiring a pathological connotation, these pseudo-items should also be perceived differently when presented alongside a label. An even more fundamental effect of such “pseudo-labels” was reported in an earlier vignette-based study: Matsunaga and Kitamura [32] found negative labeling effects even when case vignettes describing major depression were mislabeled as schizophrenia, and vice versa. We refrained from actually mislabeling symptoms with an entirely different diagnosis because we presumed that a substantial proportion of the psychology students who made up our sample would likely recognize these labels as inaccurate.

A full list of the items can be found in the appendix. The order in which the items were presented was fully randomized for each participant.

Participants were instructed to rate the behavior described in each item on a number of semantic differential scales. In the labeling condition, the instruction emphasized that these behaviors were “derived from official criteria in diagnosing mental illnesses, although not verbatim”. In the “no label” condition, no such background information on the items was given. Participants were instructed to answer as spontaneously as possible and reminded that their answers were anonymous. A sliding scale, ranging from 0 to 100, was used in order to encourage making an effort to rate each item individually rather than repeatedly choosing a middle ground, as this was deemed a likely risk in a repetitive survey on a topic with presumed social desirability in a sample that participated for course credit. If participants skipped an item by leaving the slider in the middle, they were asked to

go back and rate the item before they could proceed further in the study.

The semantic differentials on which every behavior was rated were as follows: not normal – normal; self-inflicted (“selbstverschuldet”) – not self-inflicted; rare – frequent; untreatable – treatable; affects me (“betrifft mich selbst”) – does not affect me; unacceptable – acceptable; person is in need of help – person is independent (“selbstständig” - capable of acting on their own). The instructions further clarified the intended meaning of these categories as follows. A behavior was to be rated as “self-inflicted” if it was deemed to be the person's own fault, rather than the result of circumstances beyond their control. “Treatable” referred to the viability of successful therapeutic measures. “Unacceptable” behaviors were meant to be those that a person would not tolerate in their personal environment. “Affects me” referred to whether the participants sometimes felt or behaved in this way as well. Whether the person was “in need of help” was to be judged by how much the behavior made an autonomous, independent life difficult. “Rare – frequent” as well as “normal – not normal” were deemed self-explanatory.

These categories were partly chosen to reflect typical contents of mental illness stigma [3]: personal guilt (“self-inflicted”), social distance (“acceptable”, “normal”, “affects me”), and a tendency to patronize people with mental illness because as helpless (“in need of help”). “Not treatable – treatable” was included to examine whether a possible labeling effect leads to a perception of symptoms as more chronic and unchangeable. Additionally, the categories of “self-inflicted”, “rare” and “(not) treatable” were chosen in part based on Feldman and Crandall's findings [31] that these judgments are typical for highly stigmatized mental illnesses. The stereotype of mentally ill people being dangerous, which is central for stigmatization as well according to Feldman and Crandall's findings [31], was not included because most of the symptoms included could not reasonably be perceived as threatening or dangerous to others.

The use of only these seven rating categories rather than longer scales with established psychometric properties was a necessary consequence of studying labeling effects at the item level – i. e., requiring participants to completing an established scale to measure social distance on every single item was not feasible.

In the label condition, the corresponding label was named in parentheses after each item (including the pseudo-items), as well as the ICD-10 code¹⁵ to emphasize the diagnostic nature of the labels. The instruction further clarified that these behaviors represented symptoms used to diagnose mental disorders. The no label condition did not mention mental illness or diagnostic systems. At the end of the questionnaire, participants in the label condition were debriefed to clarify that the pseudo-

¹⁵ Although the DSM was used to generate the items, participants in this German study were assumed to be more familiar with the ICD-10 codes.

items did not actually represent diagnostic criteria.

A total of 88 students enrolled in psychology at Saarland University participated in the study. By coincidence, the random allocation to the two labeling conditions resulted in two equally sized groups of 44 participants each. Of those, 74 (84%) were female, with one participant identifying as diverse and the rest as male. Participants were between 18 and 30 years old (mean = 21.7, SD = 2.8). On average, they had been studying psychology for 3.4 semesters (one and a half years). The two randomly allocated labeling groups did not differ significantly in any of these variables. Out of the 88 participants, 34 said they had at any point been diagnosed with some form of depression, 10 said so for anorexia, 7 for social anxiety, and one person each said so for schizophrenia and autism spectrum disorders. Overall, 39 out of 88 participants (44%) reported one or more of these diagnoses (partial overlap explains why the numbers do not add up to more than 39). These proportions, especially for depression, are higher than typical prevalences among the general population in Europe [33-35].

Results

Descriptive statistics and intercorrelations between the rating scales, averaged across the 24 symptoms, are presented in Table 1.

Reliability coefficients (Cronbach's Alpha) were computed for each of the rating categories for all 24 items, i. e., across diagnoses and including the pseudo-items. Values ranged from 0.88 for “affects me – does not affect me” to 0.96 for “not acceptable – acceptable”, showing that respondents' ratings of behaviors were relatively stable across these heterogeneous behaviors and even diagnoses. In other words, interindividual rating tendencies that lead participants to generally rate most behaviors as similarly “acceptable” etc. were much more prominent than intraindividual variance in ratings depending on the specific items or diagnoses.

Table 1

Means, Standard Deviations, and Intercorrelations for Rating Averages across Items

rating	<i>M</i>	<i>SD</i>	1	2	3	4	5	6
not normal - normal	38.53	12.04						
self-inflicted – not self-inflicted	60.68	16.77	-.20					
			[-.39, .01]					
rare - frequent	54.27	12.50	.39**	-.06				

			[.20, .56]	[-.27, .15]				
not treatable - treatable	74.43	11.44	-.14 [-.34, .08]	.11 [-.10, .31]	.16 [-.05, .36]			
affects me – does not affect me	31.69	14.23	.44** [.26, .60]	-.19 [-.38, .02]	.34** [.14, .51]	-.19 [-.39, .02]		
Not acceptable - acceptable	62.01	18.29	.12 [-.09, .32]	.39** [.20, .56]	.35** [.16, .52]	.31** [.11, .49]	.01 [-.20, .22]	
person is in need of help – person is independent	41.56	15.44	.35** [.15, .52]	-.06 [-.26, .15]	.27* [.07, .46]	-.15 [-.35, .06]	.28** [.07, .46]	.24* [.03, .43]

Note. Ratings can range from 0 – 100. M = mean, SD = standard deviation. 95% confidence intervals given in brackets.
* = $p < .05$. ** = $p < .01$.

Table 2 presents fit statistics from confirmatory factor analyses testing how well a model that subsumed the individual items under latent factors reflecting the five diagnoses used in the study fit the rating data. These were conducted separately for each rating category as well as for the two labeling conditions. Results show that the fit statistics generally fall short of conventional thresholds for a good model fit. This may partially result from the sample size, which was far below what is recommended for factor analysis. However, the goal of the current analysis was not to rigorously test the factor structure of the items, but to compare the model fit between the labeling conditions. In this regard, Table 2 shows that across the different rating categories, the fit statistics mostly indicate a better fit in the “label” than in the “no label” condition. In other words, the presence of a diagnostic label elicited higher intercorrelations among symptoms (including the pseudo-items) that shared the same label compared to when each symptom was presented without any label.

Table 2

Fit Statistics from Confirmatory Factor Analyses Grouped by Rating Category and Labeling condition

rating	labeling condition	CFI	RMSEA	BIC
not normal - normal	label	0.644	0.134	9170.649
	no label	0.536	0.138	9384.570

self-inflicted – not self-inflicted	label	0.759	0.145	9018.991
	no label	0.607	0.147	8976.425
rare - frequent	label	0.697	0.119	8713.773
	no label	0.697	0.131	8254900
not treatable - treatable	label	0.658	0.156	8277.304
	no label	0.551	0.156	8763.682
affects me – does not affect me	label	0.709	0.117	9271.786
	no label	0.612	0.117	9779.039
not acceptable - acceptable	label	0.776	0.137	9213.555
	no label	0.731	0.158	8880.057
person is in need of help – person is independent	label	0.749	0.125	8951.910
	no label	0.727	0.115	9254.761

Note. CFI = comparative fit index, RMSEA = root mean squared error of approximation, BIC = Bayesian information criterion. Better fit indicated by higher values in CFI and lower values in BIC.

To analyse which variables influenced the perception of individual behaviors, we conducted mixed linear regressions predicting each of the seven different ratings separately. A random intercept was included to account for the nested nature of the data (repeated measures within participants).

Additionally, the diagnosis an item pertained to, as well as the specific item itself, were also modeled as random intercepts in order to test how much of the heterogeneity in the data could be attributed to the particular behaviors that were chosen from the wider range of possible symptoms and psychiatric diagnoses. Random slopes were not included in the final analysis models because they did not improve model fit, explaining almost no variance in the results. The labeling condition (label vs. no label), the nature of the item (true symptom or pseudo-item) and a variable encoding whether or not the rater reported being afflicted with the diagnosis in question themselves, either currently or in the past, were modeled as fixed effects, along with the interactions between the labeling condition and either of the other two fixed effects. These interactions tested the possibility that the extent of labeling effects may depend on the nature of the item or the rater's own affliction with the diagnosis in question.

Results of the regression models for each of the seven outcome categories are presented in Tables 3 (fixed effects) and 4 (random effects). As the core finding of the study, the experimental condition of labeled vs. unlabeled symptoms was a significant predictor of three of the seven rating categories. In particular, participants in the label condition were more likely than those in the no label condition to rate symptoms as less common, less typical of themselves and less self-inflicted / blameworthy.

Additionally, participants who had at any point been diagnosed with the mental illness relevant to an item on average rated the behavior in question as more normal and acceptable, as more typical for themselves and as less indicative of a person being in need of external help. Besides this main effect on ratings independent of labeling condition, the diagnostic history of the respondent

furthermore interacted significantly with the labeling condition in most of the rating categories, in all cases suggesting that the presence of a label elicited more positive ratings among people currently or previously afflicted with the diagnosis in question relative to the other participants. Although average ratings of pseudo-items in some cases differed significantly from average ratings of genuine symptoms, the veracity of the item did not interact with labeling effects in any of the rating categories, suggesting that labeling effects are similar for both genuine and pseudo-items. The control variables of age and sex did not impact any of the rating averages significantly.

Table 3

Fixed Effects from Linear Models with Ratings of Behaviors as Outcomes

Rating	Effect	Estimate	SE	95% CI		p
				lower	upper	
not normal -						
normal	Intercept	17.072	12.680	-7.795	41.939	
	age	0.859	0.455	-0.033	1.751	0.062
	sex ^a	2.228	3.096	-3.844	8.300	0.474
	labeling condition ^b	-3.254	2.587	-8.327	1.820	0.212
	pseudo-item ^c	12.249	5.770	0.933	23.564	0.047
	own diagnosis ^d	-6.373	1.921	-10.140	-2.605	<0.001
	labeling condition * pseudo-item	-2.686	1.889	-6.389	1.018	0.155
	labeling condition * diagnosis	5.762	2.567	0.728	10.796	0.025
self-inflicted –						
not self-inflicted	Intercept	32.392	14.653	3.656	61.127	
	age	1.148	0.612	-0.051	2.348	0.064
	sex ^a	1.832	4.165	-6.336	10.001	0.661
	labeling condition ^b	10.409	3.433	3.677	17.141	0.003
	pseudo-item ^c	-10.559	3.862	-18.133	-2.985	0.013
	own diagnosis ^d	-0.922	1.914	-4.677	2.832	0.630
	labeling condition * pseudo-item	2.752	1.866	-0.907	6.410	0.140
	labeling condition * diagnosis	1.605	2.557	-3.409	6.620	0.530
rare - frequent	Intercept	55.175	13.127	29.431	80.918	
	age	0.214	0.474	-0.716	1.143	0.653
	sex ^a	-3.158	3.228	-9.490	3.173	0.331
	labeling condition ^b	-7.236	2.687	-12.506	-1.966	0.008
	pseudo-item ^c	8.417	3.450	1.651	15.184	0.024
	own diagnosis ^d	-2.197	1.819	-5.764	1.369	0.227
	labeling condition * pseudo-item	-2.561	1.748	-5.989	0.868	0.143
	labeling condition * diagnosis	5.387	2.405	0.671	10.103	0.025
not treatable -						

treatable		Intercept	82.190	10.555	61.491	102.890	
		age	-0.210	0.444	-1.081	0.660	0.637
		sex ^a	-2.232	3.022	-8.159	3.696	0.462
		labeling condition ^b	1.449	2.503	-3.459	6.356	0.564
		pseudo-item ^c	-2.575	1.220	-4.968	-0.183	0.040
		own diagnosis ^d	-2.828	1.570	-5.907	0.252	0.072
		labeling condition * pseudo-item	1.066	1.537	-1.948	4.080	0.488
		labeling condition * diagnosis	4.260	2.100	0.141	8.378	0.043
affects me – does							
not affect me		Intercept	54.379	13.035	28.816	79.942	
		age	-0.600	0.496	-1.571	0.372	0.230
		sex ^a	-6.681	3.373	-13.294	-0.067	0.051
		labeling condition ^b	-8.368	2.843	-13.944	-2.792	0.004
		pseudo-item ^c	2.706	5.423	-7.930	13.341	0.623
		own diagnosis ^d	10.272	2.382	5.602	14.943	<0.001
		labeling condition * pseudo-item	-0.040	2.360	-4.669	4.588	0.986
		labeling condition * diagnosis	7.510	3.189	1.256	13.764	0.019
not acceptable -							
acceptable		Intercept	51.242	16.829	18.238	84.246	
		age	0.604	0.717	-0.801	2.009	0.402
		sex ^a	-0.163	4.880	-9.734	9.408	0.973
		labeling condition ^b	2.255	4.003	-5.594	10.104	0.575
		pseudo-item ^c	-10.188	5.699	-21.365	0.989	0.087
		own diagnosis ^d	-3.853	1.898	-7.575	-0.131	0.043
		labeling condition * pseudo-item	2.307	1.834	-1.290	5.903	0.209
		labeling condition * diagnosis	7.306	2.530	2.344	12.268	0.004
person is in need							
of help – person							
is independent		Intercept	31.280	14.830	2.197	60.363	
		age	0.813	0.584	-0.331	1.958	0.167
		sex ^a	-5.386	3.975	-13.182	2.410	0.179
		labeling condition ^b	-6.248	3.288	-12.696	0.200	0.061
		pseudo-item ^c	14.546	4.089	6.527	22.565	0.002
		own diagnosis ^d	-4.203	2.024	-8.171	-0.234	0.038
		labeling condition * pseudo-item	-3.788	1.966	-7.644	0.067	0.054
		labeling condition * diagnosis	3.350	2.695	-1.936	8.635	0.214

Note. Ratings ranged from 0 – 100. ^a1 = female, 2 = male. ^b0 = no diagnostic label mentioned, 1 = diagnostic label mentioned. ^c0 = genuine diagnostic symptom, 1 = pseudo-symptom (see Methods section). ^d1 = participant reported having at any point been diagnosed with the mental illness relevant for the specific item, 0 = participant did not report such a diagnosis. SE = standard error, CI = confidence interval. Effects that are significant at $p < 0.05$ are printed in bold.

Table 4*Random Effects (Intercepts) from Models Reported in Table 3*

Rating	Effect	Proportion of variance explained
not normal - normal	diagnosis ^a	25.1%
	specific item / symptom	17.2%
	participant	14.9%
self-inflicted – not self-inflicted	diagnosis ^a	7.6%
	specific item / symptom	8.6%
	participant	34.0%
rare - frequent	diagnosis ^a	33.5%
	specific item / symptom	6.4%
	participant	18.9%
not treatable - treatable	diagnosis ^a	7.3%
	specific item / symptom	0.4%
	participant	31.7%
affects me – does not affect me	diagnosis ^a	16.0%
	specific item / symptom	12.4%
	participant	14.5%
not acceptable – acceptable	diagnosis ^a	0.0%
	specific item / symptom	17.3%
	participant	41.0%
person is in need of help – person is independent	diagnosis ^a	20.0%
	specific item / symptom	8.1%
	participant	25.6%

Note. ^aDiagnosis associated with the behavior described in the item, including pseudo-symptoms (pseudo-items). Remaining variance is residual variance.

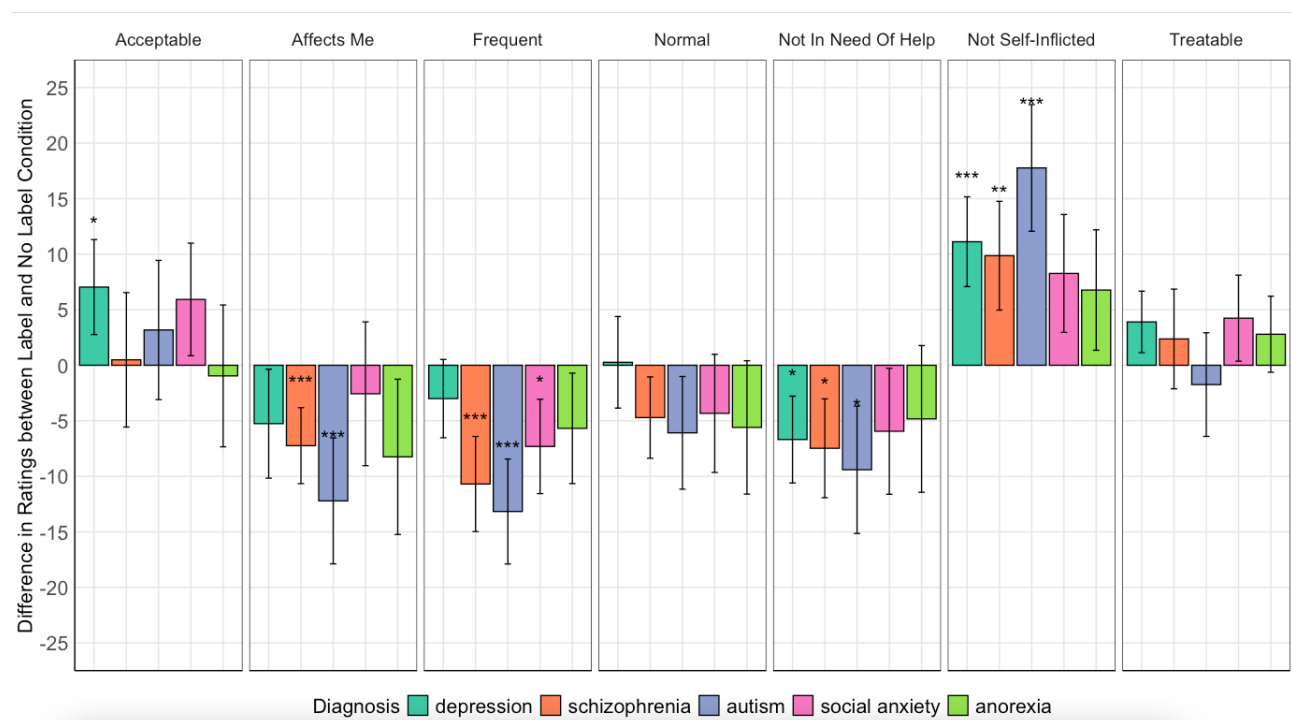
As reported in Table 4, the heterogeneity introduced by the different diagnoses and symptoms included in the survey, as well as response tendencies by individual participants, explained modest to substantial amounts of variance in the different rating categories. In line with the very high Cronbach's alpha coefficients reported earlier for ratings across items within rating categories, Table 4 shows that the participant random effects explained between 14.5% (“affects me – does not affect me”) and 41.0% (“unacceptable – acceptable”) of the variance in responses. Ratings also generally differed by the diagnosis an item pertained to. These effects were most pronounced for the rating category of “rare – frequent”, where about a third of the variance in responses was explained by this component. This may reflect the fact that some diagnoses used in the study, and by extension their symptoms, are objectively more prevalent than others. However, the random effect modeling heterogeneity in diagnoses notably did not explain variance in acceptableness even though some of the diagnoses used in the current study are typically associated with higher stigmatisation than others [31]. Besides the diagnoses an item belonged to, systematically different responses to individual items are also evident from Table 4. These were generally the smallest

effects, explaining at most 17.3% (“unacceptable – acceptable”) of the variance.

Figure 1 illustrates the labelling effects grouped by the diagnosis the symptom pertained to (or was claimed to pertain to in the study, in case of the pseudo-items) for each of the rating categories. Modeling the individual diagnoses as random effects implied that the regression models did not test specific hypotheses about which diagnoses and items influenced ratings in which direction. In Figure 1, the alpha level for significant differences shown in the graphs was corrected by a factor of 35 in order to account for the testing of five effects across seven rating scales, as there were no a priori hypotheses about specific differences. In general, numerical differences between the label and no label conditions almost uniformly point in the same direction for each rating, regardless of diagnosis. However, significant differences appear to be driven by ratings of symptoms of schizophrenia and autism more so than other diagnoses, although the lack of significant random slopes in the model makes this a very tentative conclusion that needs to be tested by further research.

Figure 1

Labeling Effects by Diagnosis and Rating Category



Note. Mean differences in absolute ratings between the label and no label conditions across different diagnoses. Since ratings can range from 0-100, differences can accordingly range from -100 to +100. Positive values indicate higher ratings of labeled versus unlabeled symptoms in the direction of the attribute specified at the top. Error bars show 95% confidence intervals. Asterisks denote p-values after applying Bonferroni corrections for multiple testing, i.e.: * $p < .05 / 35$, ** $p < .01 / 35$, *** $p < .001 / 35$.

Discussion

The current study presents a novel approach towards testing labeling effects on the perception of people with mental illness. Participants rated specific behaviors typical for five different diagnoses, as well as pseudo-items that did not reflect genuine diagnostic criteria, on seven dimensions.

Results show that mentioning a diagnostic label versus omitting it was associated with small, but significant rating differences for some of the rating categories. Labeled behaviors were perceived as less common, less typical of the raters themselves and less self-inflicted (not the person's own fault).

Notably, these labeling effects were mostly independent of other boundary conditions. Firstly, they were present across both “true” symptoms and pseudo-items, which described similar but non-pathological behavior. This lack of an interaction between the veracity of a symptom and the presence or absence of a label may suggest that the perception of behaviors that are not actually indicative of a mental illness is just as affected by (mis)labeling them as is the case for actual symptoms, an effect similar to what Matsunaga et al. [25] had found in a study using case vignettes. Secondly, descriptive results as well as the fact that introducing random slopes that would allow labeling effects to differ by diagnoses and items did not improve model fit both suggest that the effects were mostly uniform.

Notably, on five of the seven rating dimensions, labeling effects interacted significantly with the rater's own present or past affliction with the diagnosis in question. In each of these cases, the presence of a label elicited ratings in the direction of more positive perceptions (e.g., more acceptable, more common, more autonomous) among respondents who did report such a present or past diagnosis compared to those who did not. Self-stigma is an entirely different and complex field of study within the mental illness stigma literature [36], and these interaction effects suggest that research on labeling effects needs to take this issue into account as well. Further research in this regard is needed, especially because for four of the five diagnoses used in the current study, only a small minority of the sample reported having had this illness at any point.

The hypothesis that the use of diagnostic labels may be aggravating or at least facilitating mental illness stigma has a long history in practice and research, but there is still no clear consensus on how and when such labels influence the way people with mental illnesses are perceived and judged. The so far very mixed findings [23] may partially stem from different contexts, paradigms and wordings used in any given individual study, as there is no “gold standard” approach to measuring such effects. The goal of introducing yet another method in the current study was not to claim that this approach should replace the vignette paradigm, but to test whether labeling effects would still be measurable with as few context clues as possible. Results show that in some aspects, they do still

have an effect.

In line with the heterogeneity of previous findings, the labeling effects found in the current study vary by category. Ratings of how acceptable, normal, treatable or indicative of helplessness behaviors are did not robustly differ between labeling conditions in the current study. Out of the three robust labeling effects that were found, one is arguably positive - symptoms were on average perceived to be less of the person's own fault. It is noteworthy that other research suggests that this does not necessarily translate to lower stigma: focusing (solely) on biological explanation mechanisms of mental illness in educating the public has repeatedly been found to backfire by leading to more stigmatization despite also reducing impressions of blameworthiness [37, 38]. This may be because attributing mental illness to root causes outside a person's control not only reduces blame, but also the perceived chances of recovery [37]. However, such a pattern was not present in the current study as ratings of blameworthiness and of treatability – a similar concept to chances of recovery – were not significantly correlated.

Further research is needed to elucidate how labels impact perceptions of mental illness, and under which conditions specifically. For example, these processes might depend heavily on specific characteristics of the rater. Most studies on labeling effects do not control for whether the raters were themselves afflicted by a given diagnosis (see O'Connor et al., 23), which makes the inclusion of this factor in the current study valuable. However, numerous other factors that shape the impact of diagnostic labels are possible and worth investigating. Previous knowledge about the topic, previous contact with people with a certain diagnosis [39], as well as general personality characteristics might influence reactions to labeled and unlabeled behavior.

In a secondary finding, we found that the when a label was present, items belonging to the same diagnosis received more consistent ratings than unlabeled items, as indicated by improved model fit in confirmatory factor analyses. This finding supports the idea that mentioning a label leads to more generalized, stereotypical perceptions: with the label present, participants rate the behaviors with the same label more uniformly. In other words, participants apparently used their general opinion about, for example, schizophrenia to guide their ratings of specific behaviors, either consciously or unconsciously.

The current findings clearly cannot be taken to mean that using diagnostic labels in psychology is always harmful and should be abolished, or that it is ill-advised for people with mental illnesses to open up about their condition to others. Instead, studies such as the current one can attempt to clarify what consequences the use of a diagnostic label may have under which circumstances – including positive consequences. The current findings might suggest, for instance, that labeling effects could be helpful in situations in which an individual *wants* their condition to be perceived as

unusual. For example, consciously using the label of “social anxiety disorder” might serve to emphasize that a person's fear of an upcoming social situation should be taken seriously rather than being mistaken for everyday shyness. In other situations, such as a job interview, consequences such as being perceived as uncommon or less relatable may be clearly unwelcome and using a label may thus do more harm than good.

Limitations

The scope of the current study is limited by several factors: first of all, the results reported here are specific to the sample of mostly female undergraduate psychology students (which was further narrowed down by self-selection bias, as participation in the study was not mandatory for course credit). All of them were at most 30 years old as well, even though the study was not explicitly age-restricted. It stands to reason that this group does not represent the general population regarding factors such as overall stigma towards mental illness or prior familiarity with the topic.

Previous studies do show that stigmatizing attitudes and behavior are prevalent even among psychology students or even professionals in psychology and medicine [40 - 43], especially students at the bachelor rather than master level, as in the current sample [44]. Some studies even find no difference at all in stigmatizing attitudes between the general population and mental health professionals such as psychiatrists [45-46]. Accordingly, it is not unusual to investigate labeling effects among psychology students: this was the case for five of the 22 studies included in the meta-analysis [19]. In addition, ratings in the current study were mostly not obviously stigmatizing or degrading, and participants' ratings were generally well distributed (i.e., not subject to floor or ceiling effects, as might be the case in this group when measuring agreement with overtly stigmatizing statements). The fact that the mere presence of a diagnostic label still reframed participants' perceptions of behaviors even among a group that likely thinks of mental illness in a less categorical way than the general population arguably demonstrates the fundamental nature of labeling effects. When and how diagnostic labels should be used to communicate about mental illness is a relevant question not only in everyday interactions, but also in therapeutic settings such as psychotherapy or psychiatry, which is in fact one of the most frequent contexts where people with mental illness experience stigmatization [47]. Still, the specific sample in the current study naturally limits the generalizability of the current results, and a replication of the approach presented here in the general population would be desirable.

We did not explore to what extent participants in the “no label” condition may still internally have ascribed a label to the unlabeled behaviors. However, the uniformity of labeling effects, irrespective of the participant's own diagnostic history, suggests that labeling effects are constant across levels of previous knowledge. Further research is necessary to more directly explore this.

The items and diagnoses used in the current study were not previously established materials and represent a necessarily subjective and somewhat arbitrary selection among the very diverse and wide field of psychiatric diagnoses. Modeling the heterogeneity introduced by these choices as random effects in the regression models accounts for this “random” selection from a larger possible pool of items. This analysis revealed only modest influences on the variability of the ratings. Similarly, high internal consistencies across items within participants also suggested that the variability in symptoms and diagnoses only had a minor impact on ratings in the current study. Furthermore, the scope of outcomes in the current study was necessarily limited. Single-item measures such as the ones used here can work reasonably well for simple judgments [48], but they are still suboptimal, and future studies might additionally want to test for other kinds of judgments, such as the person's perceived competence and warmth [49].

Finally, the linear mixed models tested a total of 35 different coefficients. Results should therefore be interpreted with caution with regards to the possibility of spurious results due to multiple testing. However, the large number of significant and theoretically plausible effects that were found overall suggests that most of the conclusions from this study are robust. This is especially true for the labeling effects as the core research question of the study, as the three significant labeling effects that were found were significant at $p < 0.01$, not just $p < 0.05$.

Conclusion

The current study presented specific behaviors that exemplified symptoms, or pseudo-symptoms, of five different psychiatric diagnoses to a sample of psychology undergraduates. Results show that when a diagnostic label was mentioned after each of these behaviors, participants rated the behaviors on average as less typical for themselves and less common. The fictitious person showing this behavior was also rated as less blameworthy. Ratings of how normal, acceptable, treatable or indicative of helplessness the behaviors are did not differ between labeling conditions.

These results were independent of whether the behavior represented a genuine diagnostic criterion, and more generally, appeared mostly uniform across various contexts, with the exception of differing slightly among respondents who reported having experienced the mental illness in question themselves. Further research should investigate other boundary conditions.

These findings demonstrate that labeling effects partially persist at the level of specific behaviors and are therefore not simply method artifacts of the standard approach, which is to present a case vignette describing a specific person with a mental illness in a summary fashion. Similar to the findings by Martinez et al. [21], labels partly led to more positive rather than negative ratings in the form of reduced blame or guilt. However, it is also noteworthy that the two rating categories that

most resembled the standard measure of mental illness stigma – social distance – did not exhibit significant labeling effects: Behaviors were not rated as less (or more) acceptable or normal with a label than without a label. This underlines the necessity to study labeling effects beyond case vignettes, as different processes might influence outside perceptions of behavior when it is presented at a specific symptomatic level than in the broader context of a case vignette.

While the current results need to be interpreted in the context of the specific sample of psychology students and may differ when replicated among the general population, they do suggest that labeling effects exist across different mental illnesses and that they are not specific to the case vignette approach. Confirmatory factor analyses in the current study offer tentative evidence that this may partly be because the presence of a label leads to more similar ratings within the group of behaviors that are labeled the same, compared to when these behaviors come without a label.

For people with a mental illness, the question of whether and when to disclose their diagnosis often has high-stakes implications: will it backfire and lead to more stigmatizing attitudes and possibly discriminating behavior, or will it improve their quality of life through more acceptance and understanding? In general, the benefits often end up outweighing the risks, but it remains a difficult dilemma [50]. The current study cannot provide conclusive answers to this question. The best approach may depend on the specific context. Research on thinking of mental health as a continuum rather than dichotomous distinctions between “mentally healthy” and “mentally ill” has shown promising results [13]. This is not necessarily a contradiction to findings that show possible beneficial effects of using diagnostic labels, but how these two approaches may best be used in conjunction remains an open question. In any case, it is clear that labeling effects are not yet fully understood and that further research is warranted.

Conflicts of Interest: None.

Acknowledgments: The authors thank Frank M. Spinath for providing helpful comments on earlier versions of this manuscript.

Author Contributions Statement: Both authors developed the research question and contributed to designing and implementing the study. Alexander Dings implemented the statistical model and wrote the manuscript.

Informed Consent: Informed consent was obtained from all individual participants included in the study. In particular, they gave their written consent to the data privacy regulations pertaining to the study and were informed that they could cancel their participation at any point. They were also informed that participation was anonymous.

References

1. McCulloch SP, Scrivano RM. The effectiveness of mental illness stigma-reduction interventions: A systematic meta-review of meta-analyses. *Clinical Psychology Review*. 2023 Mar 1;100:102242.
2. Fekih-Romdhane F, Jahrami H, Stambouli M, Alhuwailah A, Helmy M, Shuwiekh HA, et al. Cross-cultural comparison of mental illness stigma and help-seeking attitudes: a multinational population-based study from 16 Arab countries and 10,036 individuals. *Social Psychiatry and Psychiatric Epidemiology*. 2023 Apr;58(4):641-56.
3. Rüsch N, Angermeyer MC, Corrigan PW. Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma. *Eur Psychiatry*. 2005;20(8):529-39.
4. Almomen ZA, Alqahtani AH, Alafghani LA, Alfaraj AF, Alkhalifah GS, Jalalah NH, et al. Homicide in relation to mental illness: stigma versus reality. *Cureus*. 2022 Dec;14(12).
5. Seeman N, Tang S, Brown AD, Ing A. World survey of mental illness stigma. *J Affect Disord*. 2016;190:115-21.
6. Antunes A, Silva M, Azeredo-Lopes S, Cardoso G, Caldas-de-Almeida JM. Perceived stigma and discrimination among persons with mood and anxiety disorders: Results from the WHO World Mental Health Survey Portugal. *The European Journal of Psychiatry*. 2022 Oct 1;36(4):280-7.
7. Lasalvia A, Zoppi S, Van Bortel T, Bonetto C, Cristofalo D, Wahlbeck K, et al. Global pattern of experienced and anticipated discrimination reported by people with major depressive disorder: a cross-sectional survey. *Lancet*. 2013;381(9860):55-62.
8. Pahwa R, Fulginiti A, Brekke JS, Rice E. Mental illness disclosure decision making. *Am J Orthopsychiatry*. 2017;87(5):575.
9. Mickelberg AJ, Walker B, Ecker UK, Fay N. Helpful or harmful? The effect of a diagnostic label and its later retraction on person impressions. *Acta Psychologica*. 2024 Aug 1;248:104420.
10. Pheister M, Peters RM, Wrzosek MI. The impact of mental illness disclosure in applying for residency. *Acad Psychiatry*. 2020;44:554-61.
11. Dewa CS, Van Weeghel J, Joosen MC, Gronholm PC, Brouwers EP. Workers' decisions to disclose a mental health issue to managers and the consequences. *Frontiers in Psychiatry*. 2021 Mar 26;12:631032.
12. Jones AM. Disclosure of mental illness in the workplace: A literature review. *Am J Psychiatr Rehabil*. 2011;14(3):212-29.
13. Peter LJ, Schindler S, Sander C, Schmidt S, Muehlan H, McLaren T, et al. Continuum beliefs and mental illness stigma: a systematic review and meta-analysis of correlation and

- intervention studies. *Psychol Med*. 2021;51(5):716-26.
14. Sapir E. The status of linguistics as a science. *Language*. 1929;5(4):207-14.
 15. Masland SR, Null KE. Effects of diagnostic label construction and gender on stigma about borderline personality disorder. *Stigma and Health*. 2022 Feb;7(1):89.
 16. Granello DH, Gibbs TA. The power of language and labels: “the mentally ill” versus “people with mental illnesses”. *J Couns Dev*. 2016;94(1):31-40.
 17. Fox AB, Vogt D, Boyd JE, Earnshaw VA, Janio EA, Davis K, et al. Mental illness, problem, disorder, distress: does terminology matter when measuring stigma?. *Stigma and Health*. 2021 Nov;6(4):419.
 18. Scheff T. *Being mentally ill: A sociological theory*. Chicago: Aldine; 1966.
 19. Link BG. Understanding labeling effects in the area of mental disorders: An assessment of the effects of expectations of rejection. *Am Sociol Rev*. 1987;96–112.
 20. Altmann B, Fleischer K, Tse J, Haslam N. Effects of diagnostic labels on perceptions of marginal cases of mental ill-health. *PLOS Mental Health*. 2024 Aug 28;1(3):e0000096.
 21. Link BG, Cullen FT, Struening E, Shrout PE, Dohrenwend BP. A modified labeling theory approach to mental disorders: An empirical assessment. *Am Sociol Rev*. 1989;400-23.
 22. Imhoff R. Zeroing in on the effect of the schizophrenia label on stigmatizing attitudes: a large-scale study. *Schizophr Bull*. 2016;42(2):456-63.
 23. O’Connor C, Brassil M, O’Sullivan S, Seery C, Nearchou F. How does diagnostic labeling affect social responses to people with mental illness? A systematic review of experimental studies using vignette-based designs. *J Ment Health*. 2022;31(1):115-30.
 24. Sims R, Michaleff ZA, Glasziou P, Thomas R. Consequences of a diagnostic label: a systematic scoping review and thematic framework. *Frontiers in Public Health*. 2021 Dec 22;9:725877.
 25. Franz DJ, Richter T, Lenhard W, Marx P, Stein R, Ratz C. The influence of diagnostic labels on the evaluation of students: A multilevel meta-analysis. *Educational Psychology Review*. 2023 Mar;35(1):17.
 26. Angermeyer MC, Matschinger H. The stigma of mental illness: effects of labelling on public attitudes towards people with mental disorder. *Acta Psychiatr Scand*. 2003;108(4):304-9.
 27. Manago B, Mize TD. What Is the Effect of an Inferred Mental Illness Label on Stigma? Theoretical and Empirical Challenges. *Socius*. 2024 Sep;10:23780231241274237.
 28. Martinez AG, Piff PK, Mendoza-Denton R, Hinshaw SP. The power of a label: Mental illness diagnoses, ascribed humanity, and social rejection. *J Soc Clin Psychol*. 2011;30(1):1-23.
 29. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th edn. Washington, DC: American Psychiatric Association; 2013.

30. Forbes MK, Neo B, Nezami OM, Fried EI, Faure K, Michelsen B, et al. Elemental psychopathology: Distilling constituent symptoms and patterns of repetition in the diagnostic criteria of the DSM-5. *Psychol Med*. 2024;54(5):886-94.
31. Feldman DB, Crandall CS. Dimensions of mental illness stigma: What about mental illness causes social rejection? *J Soc Clin Psychol*. 2007;26(2):137-54.
32. Matsunaga A, Kitamura T. The effects of symptoms, diagnostic labels, and education in psychiatry on the stigmatization towards schizophrenia: a questionnaire survey among a lay population in Japan. *Ment Illn*. 2016;8(1):16-20.
33. Sacco R, Camilleri N, Eberhardt J, Umla-Runge K, Newbury-Birch D. A systematic review and meta-analysis on the prevalence of mental disorders among children and adolescents in Europe. *European Child & Adolescent Psychiatry*. 2024 Sep;33(9):2877-94.
34. Jacobi F, Höfler M, Siegert J, Mack S, Gerschler A, Scholl L, et al. Twelve-month prevalence, comorbidity and correlates of mental disorders in Germany: The Mental Health Module of the German Health Interview and Examination Survey for Adults (DEGS1-MH). *Int J Methods Psychiatr Res*. 2014;23(3):304-19.
35. Thom J, Bretschneider J, Kraus N, Handerer J, Jacobi F. Versorgungsepidemiologie psychischer Störungen. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*. 2019;62(2):128-39.
36. Dubreucq J, Plasse J, Franck N. Self-stigma in serious mental illness: A systematic review of frequency, correlates, and consequences. *Schizophrenia bulletin*. 2021 Sep 1;47(5):1261-87.
37. Kvaale EP, Haslam N, Gottdiener WH. The ‘side effects’ of medicalization: A meta-analytic review of how biogenetic explanations affect stigma. *Clin Psychol Rev*. 2013;33(6):782-94.
38. Ahuvia IL, Sotomayor I, Kwong K, Lam FW, Mirza A, Schleider JL. Causal beliefs about mental illness: A scoping review. *Social Science & Medicine*. 2024 Feb 15;116670.
39. Potts LC, Henderson C. Moderation by socioeconomic status of the relationship between familiarity with mental illness and stigma outcomes. *SSM-Population Health*. 2020 Aug 1;11:100611.
40. Lien YY, Lin HS, Lien YJ, Tsai CH, Wu TT, Li H, et al. Challenging mental illness stigma in healthcare professionals and students: a systematic review and network meta-analysis. *Psychology & Health*. 2021 Jun 14;36(6):669-84.
41. Valery KM, Prouteau A. Schizophrenia stigma in mental health professionals and associated factors: A systematic review. *Psychiatry Research*. 2020 Aug 1;290:113068.
42. Riffel T, Chen SP. Stigma in healthcare? Exploring the knowledge, attitudes, and behavioural responses of healthcare professionals and students toward individuals with mental illnesses.

Psychiatric Quarterly. 2020 Dec;91(4):1103-19.

43. Jauch M, Occhipinti S, O'Donovan A. The stigmatization of mental illness by mental health professionals: Scoping review and bibliometric analysis. *PLoS One*. 2023;18(1):e0280739.
44. Prankeviciene A, Zardeckaite-Matulaitiene K, Marksaityte R, Endriulaitiene A, Tillman DR, Hof DD. Social distance in Lithuanian psychology and social work students and professionals. *Soc Psychiatry Psychiatr Epidemiol*. 2018;53(8):849-57.
45. Reavley NJ, Mackinnon AJ, Morgan AJ, Jorm AF. Stigmatising attitudes towards people with mental disorders: a comparison of Australian health professionals with the general community. *Aust N Z J Psychiatry*. 2014;48(5):433-41.
46. Nordt C, Rössler W, Lauber C. Attitudes towards people with mental illness in Europe. *Br J Psychiatry*. 2016;208(3):243-50.
47. Schulze B. Stigma and mental health professionals: A review of the evidence on an intricate relationship. *Int Rev Psychiatry*. 2007;19(2):137-55.
48. Allen MS, Iliescu D, Greiff S. Single item measures in psychological science: A call to action [Editorial]. *Eur J Psychol Assess*. 2022;38(1):1-5.
49. Görzig A, Ryan LN. The different faces of mental illness stigma: Systematic variation of stereotypes, prejudice and discrimination by type of illness. *Scandinavian Journal of Psychology*. 2022 Oct;63(5):545-54.
50. Mayer L, Corrigan PW, Eisehauer D, Oexle N, Rüsch N. Attitudes towards disclosing a mental illness: impact on quality of life and recovery. *Soc Psychiatry Psychiatr Epidemiol*. 2022;57(2):363-74.

Appendix

List of behavioral items

Order of presentation was randomized in the study. Items printed in *italics* are pseudo-items (see Methods section). Items are grouped by diagnosis below for clarity of reading; in the study, each item was presented individually and not grouped by diagnosis. In the labeling condition, the diagnosis was mentioned in brackets after the item.

Items were presented in German; original wordings are available from the authors on request. Both authors independently translated the items, agreed on a common translation and used ChatGPT to verify the accuracy of the translations (this resulted in a few minor rephrasings and grammatical corrections).

Major Depression (F32.1)

Someone is frequently tired and has little energy.

Someone often reacts irritably and abandons tasks when obstacles get in their way

Someone suffers from (severe) physical pain that increasingly occurs.

Someone has difficulty concentrating and completing tasks they have started.

Someone has difficulty completing their thoughts and making decisions.

Someone expresses frequent self-doubt and often ruminates about their own mistakes.

Someone is easily unsettled by stress.

Someone blames others for their own dissatisfaction.

Schizophrenia (F20.9)

Someone feels persecuted and surveilled in everyday life.

Someone hears internal voices that judge or command them.

Someone appears disorganized and expresses incoherent or fragmented thoughts.

Someone appears to have two distinct personalities in different situations.

Autism Spectrum Disorder (F84.0)

Someone has difficulty making friends with peers and maintaining social connections.

Someone feels uncomfortable with changes in their daily life that require adjustments to their usual routine.

Someone feels the urge to always use the same items for specific tasks.

Someone constantly interrupts others and does not seem to listen properly.

Social Anxiety Disorder (F40.10)

Someone is afraid of being judged as unintelligent or unlikeable by others.

Someone frequently worries about being criticized or judged by others.

Someone always yields to others' opinions and shows no assertiveness.

Someone reacts with insecurity to negative feedback and criticism.

Anorexia Nervosa (F50.01)

Someone has a constant fear of gaining weight, leading them to drastically reduce their food intake.

Someone overestimates their own weight and fears weight gain.

Someone assesses themselves solely based on their body weight.

Someone fears being perceived as unattractive by others.