universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Unterschiede in der subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen am Beispiel des AID 3“

verfasst von / submitted by

Theresa Renz, BSc BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2018 / Vienna 2018

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 840
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Psychologie UG2002
Betreut von / Supervisor:	Univ.- Prof. i. R. Mag. Dr. Klaus Kubinger

Danksagung

An dieser Stelle möchte ich mich zunächst für die umfangreiche Betreuung meiner Masterarbeit bei Herrn Prof. Dr. Kubinger bedanken, der immer ein offenes Ohr für meine Fragen hatte und hilfreiche Ratschläge und Anregungen geben konnte.

Ein Dank geht auch an alle Teilnehmer des Masterarbeit-Seminars, die mir oft mit kreativen Einfällen und pragmatischen Ideen weitergeholfen haben.

Nicht unerwähnt bleiben sollten meine Eltern, die mich während des ganzen Studiums moralisch und finanziell unterstützt haben und ohne die dieses Studium so nicht möglich gewesen wäre.

Inhaltsverzeichnis

Theoretischer Hintergrund

Unterschiede beim adaptiven und konventionellen Testen

Das adaptive Testen, entstanden aus der Kritik am konventionellen Testen, entwickelte sich zu einer der „herausragendsten technischen Innovationen im Bereich des psychologischen Testens im letzten Jahrhundert“ (Ortner und Caspers, 2011, S. 157). Die Kritik am konventionellen Testen besteht darin, dass allen Testpersonen die gleichen Items in derselben Reihenfolge vorgegeben werden (Kubinger, 2003). Die Vermutung lag nahe, dass auf der einen Seite leistungsstarke Testpersonen viele Items zu bearbeiten haben, die ihnen sehr leicht fallen, und auf der anderen Seite leistungsschwache Testpersonen zu viele Items erhalten, deren Bearbeitung zu schwierig für sie ist. Die Wahrscheinlichkeit für eine leistungsstarke Testperson ein sehr leichtes Item und für eine leistungsschwache Testperson ein sehr schwieriges Item zu lösen, ist in beiden Fällen mit einer hohen Wahrscheinlichkeit vorhersagbar. Viele der verwendeten Items liefern daher keine neue Information über die tatsächliche Fähigkeit der Testperson und sind daher unnötig. Die grundlegende Idee beim adaptiven Testen besteht darin, dass nur noch informative Items für die Messung einer bestimmten Fähigkeit verwendet werden sollen.

Grundsätzlich kann beim adaptiven Testen zwischen *tailored testing* und *branched testing* unterschieden werden, die im Folgenden näher erläutert werden (Kubinger, 2003). Bei der ersten Variante erhält jede Testperson als nachfolgendes Item immer nur das, dass die maximale Information über die Fähigkeit der Testperson zum gegebenen Zeitpunkt liefert. Am informativsten sind jene Items, die eine Lösungswahrscheinlichkeit von .50 für eine Testperson mit einer bestimmten Fähigkeit aufweisen (Van der Linden & Glas, 2000). Also jene Items bei denen nicht vorausgesagt werden kann, ob die jeweilige Testperson in der Lage sein wird, das Item zu lösen oder nicht. Um das informativste Item auszuwählen, ist zuvor die Berechnung der Personenparameter auf Grundlage der bisherigen Testleistung nach der *item response theory* (IRT; Lord, 1980) nötig. Die Vorgabe eines Tests nach dem *tailored testing* muss aufgrund dieser Berechnungen am Computer erfolgen. Mit jedem weiteren bearbeiteten Item wird die Schätzung der Personenparameter genauer und nähert sich mehr dem tatsächlichen Wert an. Der Test wird abgebrochen, wenn zwischen zwei aufeinander folgenden Items eine tolerierbare Abweichung der

Schätzungen der Personenparameter erreicht wurde. Das bedeutet, dass Testpersonen, je nach Fähigkeit, unterschiedlich lange Tests mit unterschiedlichen Items zur Bearbeitung erhalten (Meijers & Nering, 1999). Beim *branched testing* wird Testpersonen hingegen, je nach Gesamtleistung bei der Bearbeitung einer Gruppe von Items, nach zuvor festgelegten und festverzweigten Itemauswahlstrategien die nächste Gruppe von Items vorgegeben. Durch die festgelegten Verzweigungsregeln muss die Vorgabe der Items nicht, wie beim *tailored testing,* am Computer erfolgen, sondern kann auch als Papier-Bleistift-Verfahren, unter Anwendung verschiedenster Testmaterialien, vorgegeben werden. Für eine möglichst genaue Schätzung der Personenparameter ist dabei nicht eine höhere Anzahl an Verzweigungsschritten von Vorteil, sondern an Verzweigungsmöglichkeiten pro Schritt (Kubinger & Wild, 1989).

Gegenüber dem konventionellen Testen ergeben sich beim adaptiven Testen viele Vorteile (Meijer & Nering, 1999; Zickar, Overton, Taylor, & Harms, 1999; Weiss 1982). Da nur noch informative Items verwendet werden, haben die Tests weniger Items und sind somit ökonomischer, wobei die gemessenen Personenparameter nicht ungenauer geschätzt werden. Durch die Messung am Computer ergibt sich der zusätzliche Vorteil einer sofortigen Auswertung und Beurteilung der Fähigkeiten der Testperson. Wie anhand eines Beispiels in Kubinger (2003) veranschaulicht, erreicht man mit der konventionellen Vorgabe, von beispielsweise 53 Items, die genauste Schätzung eines Fähigkeitsparameters. Bei Vorgabe von 15 Items nach dem *tailored testing*, aus dem Itempool derselben 53 Items, nimmt die Schätzgenauigkeit bei deutlicher Zeitersparnis kaum ab. Auch die Vorgabe von 15 Items nach dem *branched testing* ist nur unwesentlich ungenauer als eine konventionelle Vorgabe oder eine nach dem *tailored testing*. Zudem gibt es zahlreiche Studien, die belegen, dass es beim adaptiven und konventionellen Testen zu vergleichbaren Schätzungen des Personenparameters kommt (Legg & Buhr, 2005; Overton, Harms, Taylor, & Zickar, 1997; Vispoel, Rocklin, & Wang, 1994). Laut einer Studie von Frey und Ehmke (2007) werden beim adaptiven Testen nur 40 bis 60 Prozent der Items benötigt, um eine genauso präzise Schätzung wie beim konventionellen Testen zu erreichen. Neben diesen Studien kommen andere zu dem Schluss, dass adaptives Testen sogar zu genaueren Schätzungen der Personenparameter führt und das mit weniger Items (Thissen & Mislevy, 2000; Wainer & Eignot, 2000; Weiss & Kingsbury, 1984). Nach der Darstellung der Vorteile des adaptiven Testens sollte aber nicht unerwähnt bleiben, dass die Konstruktion eines adaptiven Tests mit einem deutlich höheren Aufwand

verbunden ist und zur genauen Messung der Fähigkeit einer Testperson zwar weniger Items, aber ein größerer Itempool von Nöten ist (Kubinger, 2003).

Aufgrund der oben genannten Vorteile des adaptiven Testens gilt es heute als das gängigste Verfahren im Bereich der Leistungsmessung (Wainer, 2000), vor allem bei einer hohen Anzahl von Testpersonen (Ortner & Caspers, 2011). So wird es heute beim Militär (Tonidandel & Quinones, 2000; Moreno, Wetzel, McBride, & Weiss, 1984), im Bildungs- (Weiss, 2004; Weiss & Kingsbury, 1984) und Gesundheitswesen (Gibbons et al., 2008; Fayers, 2007) standardgemäß eingesetzt.

Während die oben genannten Vorteile und andere technische Aspekte des adaptiven Testens, wie die Größe des Itempools, verschiedene Item-Auswahl-Verfahren, verschiedene Schätzverfahren und deren Gütekriterien, oft untersucht bzw. beschrieben wurden (Hau & Chang, 2001; Chen, Ankenmann, & Chang, 2000; Thissen & Mislevy, 2000; Van der Linden & Glass, 2000; Wainer & Eignor, 2000; Meijer & Nering, 1999; Wang & Vispoel, 1998; Veerkamp & Berger, 1997), gibt es hingegen kaum Studien über mögliche unterschiedliche psychische Auswirkungen auf die Testpersonen beim adaptiven und konventionellen Testen (Tonidandel & Quinones, 2000). Das Interesse an diesen Unterschieden und der damit möglicherweise zusammenhängenden unterschiedlichen subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen, stieg aber in den letzten Jahren an (Ortner, Weißkopf, & Koch, 2014).

Selbsteinschätzung der eigenen Leistung beim adaptiven und konventionellen Testen

Wie oben bereits erwähnt, erhält man beim adaptiven und konventionellen Testen annähernd dieselben Personenparameter (Overton et al., 1997; Legg & Buhr, 2005; Vispoel et al., 1994), aber die subjektive Einschätzung der eigenen Leistung fällt beim adaptiven und konventionellen Testen möglicherweise unterschiedlich aus. So fanden Macan, Avedon, Pease und Smith (1994) zwischen der eingeschätzten und der tatsächlichen Leistung bei einem konventionellen kognitiven Leistungstest eine signifikante Korrelation von $r = .40$, die aber nur etwa 16 Prozent der erklärten Varianz entspricht und nur als unbedeutend interpretiert werden muss. Im Gegensatz dazu fand Powell (1994) beim adaptiven Testen keinen signifikanten Zusammenhang

zwischen der eingeschätzten und der tatsächlichen Leistung der Testpersonen (r = .07). In einer weiteren Studie (Ortner, Weißkopf, & Koch, 2014), in der adaptives und konventionelles Testen direkt miteinander verglichen wurden, konnte gezeigt werden, dass Testpersonen, unabhängig von ihrem Leistungsniveau, beim adaptiven Testen nicht in der Lage waren ihre Leistungen einzuschätzen (β = –.01, t = -0.08, p = .94). Vielmehr schätzen sowohl leistungsstarke als auch leistungsschwache Testpersonen ihre Leistungen beim adaptiven Testen gleich ein. Die tatsächliche Leistung im adaptiven Test war somit kein Prädiktor für die eingeschätzte Leistung. In der konventionellen Bedingung wurde hingegen eine signifikante Steigung der tatsächlichen Leistung (β = .41, t = 3.89, p < .01), die auf einen signifikanten linearen Zusammenhang zwischen eingeschätzter und tatsächlicher Leistung hindeutete, gefunden. Die tatsächliche Leistung erklärte aber, innerhalb eines größeren Modells, in dieser Studie weniger als 4 Prozent der erklärten Varianz und muss daher als unbedeutend interpretiert werden. Die Testpersonen konnten ihre tatsächlichen Leistungen im konventionellen Test somit nur bedingt einschätzen.

Ein Grund für eine potentiell unterschiedliche Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen könnte die Art sein, wie Testpersonen üblicherweise ihre Leistungen in einem Test beurteilen und die spezielle Auswertung beim adaptiven Testen (Ortner, Weißkopf, & Gerstenberg, 2013; Tonidandel & Quinones, 2000). Beim konventionellen Testen ergibt sich die Messung der Fähigkeit einer Testperson aus der Summe an richtig beantworteten Items, wie beispielsweise auch in den meisten Schultests. Ist eine Testperson nach dem Test in der Lage einzuschätzen, wie viele Items sie gelöst hat, kann sie auch relativ gut einschätzen, wie die Beurteilung der eigenen Fähigkeiten ausfallen wird. Beim adaptiven Testen hingegen ist nicht nur die Anzahl der gelösten Items entscheidend, sondern auch welche Items korrekt gelöst wurden. Der Personenparameter ist umso höher, je mehr schwierige Items richtig beantwortet wurden. So wird eine leistungsschwache Person beim konventionellen Testen den Anteil, der von ihr richtig bearbeiteten Items, rein theoretisch, niedrig einschätzen und ihre Leistung in dem Test negativ beurteilen. Beim adaptiven Testen wird sie, unabhängig von ihren Fähigkeiten, in etwa die Hälfte der Items lösen können und daher möglicherweise den Eindruck gewinnen, dass sie, auch im Vergleich zu bisherigen Leistungen in Schultests, einen höheren Anteil an Items gelöst hat und daher annehmen, eine bessere Leistung erbracht zu haben. Leistungsstarke Personen sind es bei konventionellen Tests

hingegen gewohnt, dass sie die meisten Items lösen können. Beim adaptiven Testen werden aber auch sie nur in etwa die Hälfte der Items lösen können und dürften sich daher, vor allem im Vergleich zu bisherigen Schultests, eher negativer einschätzen bzw. unzufriedener mit ihrer Leistung sein. Während leistungsschwache und leistungsstarke Testpersonen den Anteil der gelösten Items beim konventionellen Testen sehr unterschiedlich einschätzen müssten, sollten beide diesen Anteil beim adaptiven Testen in etwa gleich hoch einschätzen, da beide Gruppen jeweils die Hälfte der Items lösen können.

Für diese Theorie sprechen die Studienergebnisse von Ortner et al. (2013), die übereinstimmend mit den oben dargestellten Forschungsergebnissen von Macan et al. (1994), Powell (1994) und Ortner et al. (2014) einen signifikant positiven Zusammenhang (r = .37, p < .01) zwischen dem eingeschätzten Prozentsatz an gelösten Items und der tatsächlichen Leistung beim konventionellen Testen fanden, was allerdings weniger als 14 Prozent erklärter Varianz entspricht und daher nicht als bedeutender Zusammenhang interpretiert werden kann. Sie konnten jedoch keinen signifikanten Zusammenhang (r = -.01) dieser beiden Variablen beim adaptiven Testen beobachten. Je besser Testpersonen ihre Leistungen in einem konventionellen Test einschätzen, desto leichter bzw. weniger anstrengend dürften sie diesen Test empfinden. Interessanterweise fanden Ortner et al. (2013) keinen signifikanten Zusammenhang zwischen der tatsächlichen Leistung und der eingeschätzten Schwierigkeit des Tests (r = -.19) beim konventionellen Testen und einen zwar signifikanten aber unbedeutenden Zusammenhang dieser beiden Variablen beim adaptiven Testen (r = .32, p < .01), der nicht viel mehr als 10 Prozent der Varianz erklärte. Je höher die Fähigkeiten der Testpersonen waren, desto schwieriger stuften sie den adaptiven Test ein. Dass leistungsschwache Teilnehmer adaptive Tests möglicherwiese als leichter und weniger anstrengend empfinden, könnte daran liegen, dass sie im Gegensatz zu leistungsstarken Teilnehmern tatsächlich leichtere Items vorgelegt bekommen und umgekehrt (Tonidandel & Quinones, 2000). Ortner et al. (2013) konnten zudem einen signifikant positiven Zusammenhang zwischen der tatsächlichen Fähigkeit und der eingeschätzten Anstrengung beim adaptiven Testen (r = .28, p < .01) finden, der aber mit einer erklärten Varianz von weniger als 8 Prozent als unbedeutender Effekt interpretiert werden muss. Beim konventionellen Testen konnten sie keinen Zusammenhang dieser beiden Variablen (r = .02) beobachten. Ebenso fanden die Autoren einen positiven Zusammenhang zwischen der

tatsächlichen Leistung und der Zufriedenheit mit der Leistung in der konventionellen Bedingung ($r = .45$, $p < .01$), der aber auch nur etwa 20 Prozent der Varianz erklären konnte. In der adaptiven Bedingung wurde dagegen ein signifikant negativer Zusammenhang gefunden ($r = -.21$, $p < .05$), d.h. je besser die Leistung der Testpersonen war, desto unzufriedener waren sie mit ihrer Leistung. An dieser Stelle sollte allerdings betont werden, dass dieser Effekt weniger als 5 Prozent der erklärten Varianz entspricht. Die Autoren nahmen an, dass die Beziehung zwischen der tatsächlicher Leistung einer Testperson und deren Zufriedenheit mit der eigenen Leistung, der empfundenen Schwierigkeit des Tests und der benötigten Anstrengung indirekt über die subjektive Einschätzung der eigenen Leistung beeinflusst wurde. Ebenso wie die subjektive Einschätzung der eigenen Leistung Einfluss darauf haben könnte, welche Attribute man einem Test zuschreibt, könnte sie auch Einfluss auf die Zufriedenheit mit der eigenen Leistung und der eigenen Motivation haben.

Tonidandel, Quinones und Adams (2002) postulierten, dass die subjektive Einschätzung der eigenen Leistung die psychischen Reaktionen (Zufriedenheit, Motivation und Ängstlichkeit) der Testpersonen auf einen adaptiven Test beeinflussen würden, was sie auch empirisch belegen konnten. Im Widerspruch zu den Ergebnissen von Ortner et al. (2013) fanden Tonidandel et al. (2002) einen positiven Zusammenhang zwischen eingeschätzter Leistung und Zufriedenheit beim adaptiven Testen. Tonidandel et al. (2002) konnten nicht nur einen positiven Zusammenhang zwischen Zufriedenheit und eingeschätzter Leistung ($r = .25$, $p < .05$), sondern auch zwischen Zufriedenheit und Anzahl an gelösten Items ($r = .15$, $p < .05$) und zwischen Zufriedenheit und tatsächlicher Leistung ($r = .20$, $p < .05$) beim adaptiven Testen finden. Ähnliche Ergebnisse fanden sie auch für den Zusammenhang zwischen Motivation und tatsächlicher Leistung ($r = .08$, $p < .05$), Anzahl an gelösten Items ($r = .18$, $p < .05$) und eingeschätzter Leistung ($r = .28$, $p < .05$). Die Autoren konnten ihr Model, wonach die eingeschätzte Leistung die Beziehung zwischen objektiver Schwierigkeit des Tests (operationalisiert durch die Einschätzung der Anzahl an gelösten Items) und der jeweiligen Variable indirekt beeinflusst, sowohl in Bezug auf die Zufriedenheit als auch auf die Motivation, bestätigen. Allerdings handelte es sich bei allen Zusammenhängen um kleine Effekte, wobei der größte von ihnen sich in weniger als 8 Prozent erklärter Varianz niederschlug.

Welchen Einfluss die Tatsache, dass alle Testpersonen leistungsunabhängig in etwa die Hälfte der Items lösen können und die Schwierigkeit der zu bearbeiteten Items von der erbrachten bisherigen Leistung der Testpersonen abhängt, auf die Motivation der Testpersonen haben, ist in der Literatur allerdings umstritten (Ortner & Caspers, 2011; Frey, Hartig, & Moosbrugger, 2009). Aufgrund der Tatsache, dass leistungsstärkere Testpersonen im Verlauf eines adaptiven Leistungstests schwierigere Items und leistungsschwache leichtere Items vorgelegt bekommen, wurde argumentiert, dass sich erstere im Vergleich zu einer konventionellen Testung weniger unterfordert und letztere weniger überfordert fühlen würden (Weiss & Betz, 1973, zitiert nach Ling, Attali, Finn, & Stone, 2017). In beiden Gruppen sollte demnach die Motivation zur Bearbeitung des adaptiven Tests, durch diese genaue Passung zwischen der Schwierigkeit der Items und der eigenen Leistung, höher sein. Das adaptive Testen müsste somit seltener in Frustration oder Langeweile enden. Eine aktuelle Studie (Martin & Lazendic, 2018) berichtete, dass vor allem SchülerInnen der neunten Klasse angaben beim adaptiven Testen motivierter zu sein, obwohl Jugendliche in diesem Alter ansonsten nur wenig Engagement für die Schule zeigen (Martin, 2007). Jedoch beobachteten die Autoren tatsächlich nur eine signifikante Interaktion zwischen Testbedingung und Schulklasse von $\beta = 0.02$ für die Motivation der SchülerInnen der neunten Klasse, die kaum bedeutend ist. Die eingeschätzte Motivation der NeuntklässlerInnen unterschied sich zwischen der konventionellen ($M = 4.22$, $SD = 1.17$) und adaptiven Bedingung ($M = 4.29$, $SD = 1.12$) auf einer siebenkategorialen Ratingskala kaum. Im Gegensatz dazu vermuteten Eggen und Verschoor (2006), dass adaptive Tests vor allem jüngere Schüler demotivieren dürfte, da sie höhere Lösungswahrscheinlichkeiten bei gewöhnlichen Schultests gewohnt seien (Bergstorm & Lunz, 1999; Bergstrom, Lunz, & Gershon, 1992). Dies führe dazu, dass sie die Tests schwieriger oder ihre eigenen Fähigkeiten niedriger einschätzen würden als bei konventionellen Tests. Die Autoren argumentierten, dass daher vor allem bei leistungsstarken Testpersonen die Motivation beim adaptiven Testen sinken müsste. Frey et al. (2009) fanden einen kleinen Haupteffekt des Faktors Adaptivität ($F(1,77) = 4.14$, $p = .045$, $\eta^2 = .051$), wobei die eingeschätzte Motivation beim konventionellen Testen ($M = 25.10$, $SD = 17.40$) größer ausfiel als beim adaptiven Testen ($M = 20.16$, $SD = 13.80$). Ihrer Ansicht nach würden die Testpersonen erkennen, dass sie nur in etwa die Hälfte der Items lösen konnten und würden sich daher schlechter als bei einem konventionellen Test zur Konzentrationsleistung

einschätzen. An dieser Stelle sollte aber nicht unerwähnt bleiben, dass die hier gewählte adaptive Form sehr speziell erscheint. Bei diesem Test entscheidet nicht die bisherige Testleistung darüber, wie schwierig die nächsten Items sein werden, sondern wie lange zehn Items gleichzeitig auf einem Bildschirm dargeboten werden. Je nach Testleistung wurden die Items so lange dargeboten, dass eine Wahrscheinlichkeit von 50 Prozent bestand, dass alle Items innerhalb der zur Verfügung stehenden Zeit gelöst werden konnten. Ling et al. (2017) versuchten die unterschiedlichen Forschungsergebnisse dadurch zu erklären, dass in den einzelnen Untersuchungen sehr unterschiedliche Tests für die Untersuchung der psychischen Reaktionen beim adaptiven und konventionellen Testen verwendet wurden. Sie entschieden sich in ihrer Untersuchung daher für einen Test, der die mathematischen Fähigkeiten untersucht und den sie als mental schwieriger als andere Leistungstests, die beispielsweise nur bestehendes Wissen abfragen, einstuften. Sie vermuteten, dass adaptives Testen bei kognitiv anspruchsvolleren Aufgaben motivierender wirken würde, konnten diese Hypothese aber nicht bestätigen. Sie fanden zwar einen kleinen signifikanten Haupteffekt der Testbedingung ($F(2, 765) = 12.29$, $p < .01$, $\eta^2 = .031$) mit einem höheren Mittelwert für die konventionelle Bedingung ($M = 3.10$, $SD = 0.05$) als für die adaptive ($M = 2.97$, $SD = 0.05$). Der gefundene Haupteffekt muss jedoch als kleiner Effekt klassifiziert werden. Als Kritik an dieser Studie sei angemerkt, dass die Autoren nur einen mathematischen Leistungstest verwendeten, um ihre Hypothese zu untersuchen, und nicht noch andere mental anspruchsvolle Tests vorlegten.

Ein weiterer Grund für die Unterschiede der subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen könnte darin liegen, dass es beim adaptiven Testen nicht die Möglichkeit gibt, Items auszulassen oder auf sie zurückzukommen, um sie zu einem späteren Zeitpunkt zu verbessern (Tonidandel & Quinones, 2000). In einer Studie führten Tonidandel und Quinones (2000) eine Befragung über die empfundene Fairness, Einstellung und Erwartung gegenüber Tests, die entweder adaptiv oder konventionell geschildert wurden, durch. Die Befragung ergab, dass die fehlende Möglichkeit Items zu überspringen und zu verbessern sowie dass nicht alle Testpersonen dieselben Items zu bearbeiten haben zu negativen Reaktionen der Testpersonen führten und dass das adaptive Format daher als unfair empfunden wurde. Nicht unerwähnt bleiben sollte hier aber, dass bei dieser Studie nur die Einstellungen erfragt, aber kein adaptiver Test durchgeführt wurde, um danach nach der empfundenen Fairness der verschiedenen

Testbedingungen zu fragen. Aber es erscheint intuitiv einleuchtend, dass eine negative Einstellung gegenüber einer Testung dazu führen könnte, dass der Test schwieriger empfunden wird, die Testpersonen selbst weniger zufrieden und motiviert sind und sie ihre Leistungen schlechter einschätzen oder sich stärker an jene Items erinnern, von denen sie wissen, dass sie von ihnen falsch bearbeitet wurden und nicht verbessert werden konnten.

Ein weiterer möglicher Grund für den Unterschied der subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen ist das bereits erwähnte Abbruchkriterium beim adaptiven Testen (Tonidandel & Quinones, 2000), dessen Anwendung in der Regel in einer kürzeren Testlänge resultiert. Vispoel et al. (1994) vermuteten darin die eigentliche Attraktivität des adaptiven Testen für Testpersonen und erklärten die Beliebtheit dieses Formats unter Testpersonen (Legg & Buhr, 1992; Moe & Johnson, 1988; Schmidt, Urry, & Gugel, 1978) durch dieses Zeitersparnis. Durch die kürzere Testlänge empfinden Testpersonen adaptive Tests eventuell als weniger anstrengend und bewerten ihre Leistungen positiver.

Die oben beschriebenen Ursachen für mögliche Unterschiede der subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen treffen möglicherweise aber mehr auf das *tailored testing* zu. Die Unterschiede zwischen *tailored* und *branched testing* könnten auch die subjektive Einschätzung der eigenen Leistung beeinflussen. Zwar werden beim *branched testing,* nach der Bearbeitung der ersten Gruppe von Items, die nächsten Items auch leistungsabhängig ausgesucht, jedoch lösen die Testpersonen bei dieser Variante des adaptiven Testens oft nicht genau die Hälfte der Items. Beim *branched testing* liegt die Wahrscheinlichkeit einer Testperson fast alle Items oder nur sehr wenig zu lösen im Bereich des Möglichen. Leistungsstarke Testpersonen können beim *branched testing*, wie beim konventionellen Testen, durchaus in der Lage sein fast alle Items zu lösen, während leistungsschwache Testpersonen möglicherweise fast keine Items lösen können. Jedoch erscheint es auch beim *branched testing* wahrscheinlich, dass Testpersonen versuchen über den Anteil an gelösten Items ihre Leistung zu beurteilen. Möglicherweise können daher sehr gute und sehr schlechte Testpersonen ihre Leistungen, da sie in der Lage waren viele schwierigere Items bzw. kaum leichtere Items zu lösen, beim *branched testing*, quasi zufällig, akkurater einschätzen. Wie beim *tailored testing* erhalten leistungsstarke Testpersonen schwierigere Items als

leistungsschwache, aber im Gegensatz zum *tailored testing* gleich eine ganze Gruppe schwierigerer Items. Das könnte auch beim *branched testing* dazu führen, dass die Tests, von leistungsstarken Testpersonen als schwieriger oder anstrengender und im Gegensatz dazu von leistungsschwachen als leichter und weniger fordernd empfunden werden. Bei dieser Variante des adaptiven Testens besteht zudem die Möglichkeit, Items, der gerade zu bearbeitenden Gruppe, zu verbessern. Die Testlänge und ausgewählten Items unterscheiden sich bei verschiedenen Teilnehmern zudem nicht in dem Ausmaß, wie das beim *tailored testing* möglich ist. Die von Tonidandel und Quinones (2000) berichtete empfundene negative Einstellung gegenüber dem adaptiven Testen wäre beim *branched testing* eventuell weniger negativ ausgefallen. Jedoch liegen keine Studien vor, die die Auswirkungen der geringen Unterschiede beim *branched* und *tailored testing* auf die subjektive Einschätzung der eigenen Leistung untersuchen haben.

Verzerrungen der subjektiven Einschätzung der eigenen Leistung und Einfluss des Alters auf die subjektive Einschätzung der eigenen Leistung

Bei der Betrachtung der möglichen Unterschiede der subjektiven Selbsteinschätzung der eigenen Leistung beim adaptiven und konventionellen Testen sollte der Einfluss allgemeiner Verzerrungen auf die subjektive Selbsteinschätzung nicht vergessen werden.

Zunächst sei die oft beobachtete Tendenz durchschnittlicher Personen genannt, sich selbst in bestimmten wünschenswerten Eigenschaften oder Fähigkeiten als überdurchschnittlich einzuschätzen. Unter anderem nahmen die meisten Menschen nicht nur an, dass sie einfühlsamer, verständnisvoller und empathischer als andere seien, sondern auch fähiger, kompetenter und talentierter (Brown, 2007; Dunning, Health, & Suls, 2004). So betrachteten sich Geschäftsmänner durchschnittlich als fähiger als andere (Larwood & Whittaker, 1977) und Fußballspieler gaben an, Fußball an sich besser zu verstehen als ihre Mannschaftskollegen (Felson, 1981).

Im Einklang mit dieser Tendenz konnte auch belegt werden, dass Personen, die in einer bestimmten Eigenschaft oder Fähigkeit als unterdurchschnittlich eingestuft wurden, sich selbst meist als durchschnittlich, wenn nicht sogar überdurchschnittlich,

beschrieben und damit noch weniger als durchschnittliche Personen in der Lage waren sich selbst einzuschätzen (Kruger & Dunning, 1999). Beispielsweise war sich der Großteil sozial inkompetenter Jungen und junger Männer dieser Inkompetenz nicht bewusst (Fagot & O'Brien, 1994; Bem & Lord, 1979). Auch leistungsschwache Studierende waren kaum in der Lage, ihre Leistungen in einer Lehrveranstaltung akkurat zu beurteilen (Moreland, Miller, & Laucka, 1981) oder einzuschätzen, welche Fragen sie in einer Prüfung falsch beantwortet hatten (Shaughnessy, 1979; Sinkavich, 1995). Außerdem schätzten auch Menschen mit Lernstörungen ihre Leistungen besser ein als Menschen ohne (Stone & May, 2002; Stone, 1997; Alvarez, Adelmann, 1986).

Kruger und Dunning (1999) konnten nicht nur durch mehrere Untersuchungen belegen, dass sich leistungsschwache Personen, deren Leistung im untersten Perzentil lagen, ihre Leistungen oder Fähigkeiten deutlich überschätzten, sondern auch, dass sich leistungsstarke Personen unterschätzten. Menschen aus dem 12. Perzentil stuften ihre Leistungen bei Tests zum logischen Denken und zu Grammatik durchschnittlich im 62. Perzentil ein. Beim logischen Denken stuften Menschen aus dem 86. Perzentil ihre Fähigkeiten als dem 68. Perzentil zugehörig ein. Ähnlich verhielt es sich bei einem Test, der die Grammatikkenntnisse überprüfte. Hier schätzten Testpersonen aus dem 89. Perzentil ihre Leistungen als im 72. Perzentil liegend ein. Die Autoren erklärten ihre Ergebnisse damit, dass leistungsschwache Personen ihre eigenen Fähigkeiten und leistungsstarke die Fähigkeiten der anderen Testpersonen überschätzen würden. Sie argumentieren weiter, dass sich inkompetente Personen nicht nur aufgrund mangelnder metakognitiver Fähigkeiten (Everson & Tobias, 1998) schlechter einschätzen können würden, sondern auch, dass diese mangelnden Fähigkeiten der Grund seien, wieso sie überhaupt inkompetent bleiben. Demnach seien sie nicht in der Lage, ihre eigenen Fehler oder Schwachstellen zu erkennen und würden ihre Fähigkeiten dadurch auch nicht verbessern können. Die Evidenz der obigen Ergebnisse konnte durch eine aktuelle Studie (Kim, Kwon, Lee, & Chiu, 2016) untermauert werden. So konnte ein positiver Zusammenhang zwischen tatsächlicher und eingeschätzter Leistung gefunden werden ($r = .45$, $p < .05$), der etwa 20 Prozent der Varianz erklärt. Betrachtete man die Quartile aber genauer, ergab sich ein differenzierteres Bild. Testpersonen deren Leistung im ersten Quartil lagen, schätzen ihre Fähigkeiten im Durchschnitt 37.85 Perzentile zu hoch ein, während Testpersonen, deren Leistung sich im vierten Quartil befanden, ihre Fähigkeiten im Durchschnitt um 17.70 Perzentile zu niedrig einschätzten.

Welche Auswirkungen diese Verzerrungen auf die subjektive Einschätzung der eigenen Leistung beim adaptiven Testen haben, wurde bisher allerdings nicht direkt untersucht. Ortner et al. (2013) beobachteten lediglich, wie oben bereits erwähnt, dass, umso schlechter die Leistung einer Testperson in der adaptiven Bedingung war, desto zufriedener war sie mit ihren Leistungen und desto leichter und weniger fordernd empfand sie den Test. In der konventionellen Bedingung fanden sie diesen Effekt nicht. Jedoch untersuchten sie die obersten und untersten Perzentile weder beim adaptiven noch beim konventionellen Testen genauer.

Nicht unbeachtet bleiben sollte aber auch, dass das Alter der Testpersonen Einfluss auf die Fähigkeit, die eigene Leistungen akkurat einschätzen zu können, haben könnte. So bewerteten Kindergartenkinder oder Kinder der ersten Schulkasse ihre Fähigkeiten meistens mit der bestmöglichen Zuschreibung (Miller, 1987; Stipek, 1981) und erst Kinder ab der dritten Schulklasse wurde die Fähigkeit die eigenen Leistungen akkurat einschätzen zu können, zugeschrieben (Licht, 1992; Nicholls, 1979; Nicholls 1978). Einer etwas aktuelleren Studie (Kaderavek, Gillam, Ukrainetz, Justice, & Eisenberg, 2004) zufolge überschätzten 85 Prozent der fünf- bis neunjährigen Kinder ihre Fähigkeiten und 43 Prozent der zehn- und elfjährigen Kinder. Bei letzteren konnte aber auch nur ein kleiner, signifikant positiver Zusammenhang zwischen tatsächlicher und eingeschätzter Leistung gemessen werden (r = .22, p = .02), was weniger als 5 Prozent der erklärten Varianz entspricht. Untersucht wurde in dieser Studie aber nur die eingeschätzte und tatsächliche Leistung bei der mündlichen Sprachproduktion unter einer konventionellen Testbedingung. Jüngere Kinder scheinen ihre Fähigkeiten aber, laut der hier beschriebenen Literatur, tendenziell mehr zu überschätzen als ältere. Da Testpersonen ihre Leistungen in einem adaptiven Test generell nur unzureichend einschätzen können (Ortner et al., 2014; Macan, 1994), wird an dieser Stelle vermutet, dass sich das Alter der Kinder nur beim konventionellen Testen in einer etwas akkurateren Selbsteinschätzung niederschlagen könnte, wobei die beobachteten Zusammenhänge zwischen tatsächlicher und eingeschätzter Leistung, wie bereits erwähnt, auch beim konventionellen Testen, schwach sind.

Fragestellungen

Adaptives und konventionelles Testen unterscheiden sich in der Durchführung und der Auswertung. Diese Unterschiede könnten, wie oben bereits ausführlich dargestellt, Auswirkungen auf die subjektive Einschätzung der eigenen Leistungen beim adaptiven und konventionellen Testen haben. Im Einklang mit dem aktuellen Forschungsstand lautet die erste Hypothese dieser Studie folgendermaßen:

Hypothese 1: Es gibt einen signifikanten Unterschied in der subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen.

Des Weiteren zeigte sich, dass leistungsstärkere Testpersonen ihre Leistungen beim konventionellen Testen höher und leistungsschwächere Testpersonen ihre Leistungen niedriger einstuften. Im Gegensatz dazu schätzten Testpersonen ihre Leistungen beim adaptiven Testen leistungsunabhängig in etwa gleich hoch ein. So wurde in verschiedenen Studien eine signifikant positive Korrelation zwischen eingeschätzter und tatsächlicher Leistung beim konventionellen Testen gefunden, allerdings spiegelten sich diese Zusammenhänge nur in unbedeutenden Effektstärken wider. Beim adaptiven Testen wurde, wahrscheinlich aufgrund der Tatsache, dass jede Testperson gleich viele Items lösen kann, kein signifikanter Zusammenhang zwischen diesen beiden Variablen beobachtet. Da in dieser Studie adaptiv im Sinne des *branched testing* getestet wird, könnte der Zusammenhang zwischen tatsächlicher und eingeschätzter Leistung etwas höher sein als bisher in der Literatur, in Bezug auf das adaptive Testen, berichtet wurde. Jedoch wird hier nicht davon ausgegangen, dass sich die geringen Unterschiede zwischen *branched* und *tailored testing* allzu deutlich in der subjektiven Einschätzung der eigenen Leistung auswirken werden und keine starken positiven Zusammenhänge zwischen der eingeschätzten und tatsächlichen Leistung gemessen werden. Darum lauten die folgenden Hypothesen:

Hypothese 2a: Es besteht kein signifikant positiver Zusammenhang in relevantem Ausmaß ($r > .70$) zwischen der tatsächlichen Leistung und der subjektiven Einschätzung der eigenen Leistung beim konventionellen Testen.

Hypothese 2b: Es besteht kein signifikanter Zusammenhang in relevantem Ausmaß ($r > .70$) zwischen der tatsächlichen Leistung und der subjektiven Einschätzung der eigenen Leistung beim adaptiven Testen.

Wie bereits oben erwähnt, konnte beobachtet werden, dass sehr leistungsschwache Testpersonen ihre Leistungen überschätzen und dass sehr leistungsstarke Testpersonen ihre Leistungen unterschätzen. Dieser Effekt (Dunning-Kruger-Effekt) wurde jedoch noch nicht direkt beim adaptiven und konventionellen Testen verglichen. Zwar wird in dieser Studie davon ausgegangen, dass sich leistungsstarke Testpersonen beim adaptiven Testen ohnehin etwas unterschätzen und leistungsschwache etwas überschätzen, dennoch wird vermutet, dass die verzerrte Einschätzung sehr guter und sehr schlechter Testpersonen beim adaptiven Testen nicht viel stärker zum Ausdruck kommt als beim konventionellen Testen. Die dritte Hypothese lässt sich, mit Vorsicht, da noch keine einschneidenden Studien dazu vorliegen, folgendermaßen formulieren:

Hypothese 3: Es gibt weder einen Unterschied im Ausmaß der Verzerrung zwischen der subjektiven Einschätzung der eigenen Leistung und der tatsächlichen Leistung bei leistungsstarken noch bei leistungsschwachen Testpersonen beim adaptiven Testen und konventionellen Testen.

Zu guter Letzt wurde dargelegt, dass angenommen wird, dass ältere Kinder ihre Leistungen, im Gegensatz zu jüngeren Kindern, akkurater einschätzen können. Allerdings konnten auch bei älteren Kindern nur unbedeutende Zusammenhänge berichtet werden. Dabei handelte es sich ausschließlich um Studien, die die Einschätzung der eigenen Leistung von Kinder in konventionellen Testbedingungen untersuchten. Es wird in dieser Studie davon ausgegangen, dass sich auch in der adaptiven Bedingung weder die älteren noch die jüngeren Kinder akkurat einschätzen können. Die folgenden Hypothesen lauten daher:

Hypothese 4a: Es gibt weder beim adaptiven noch beim konventionellen Testen einen signifikanten Zusammenhang in relevantem Ausmaß ($r > .70$) zwischen der subjektiven Einschätzung der eigenen Leistung und der tatsächlichen Leistung in der Gruppe der jüngeren Kinder.

Hypothese 4b: Es gibt weder beim adaptiven noch beim konventionellen Testen einen signifikanten Zusammenhang in relevantem Ausmaß ($r > .70$) zwischen der subjektiven Einschätzung der eigenen Leistung und der tatsächlichen Leistung in der Gruppe der älteren Kinder.

Methoden

Untersuchungsdesign

Bei der hier beschriebenen Studie handelt es sich um eine experimentelle Untersuchung. Die Testpersonen wurden zu Beginn der Untersuchung auf zwei verschiedene experimentelle Bedingungen, adaptives und konventionelles Testen, verteilt. Die Zuteilung zu den Gruppen erfolgte randomisiert durch einen Würfel. Wurde eine gerade Zahl gewürfelt, wurde die Testperson adaptiv getestet, während bei einer ungeraden Zahl eine konventionelle Testung durchgeführt wurde. Beide Versuchsbedingungen unterschieden sich nur in der Form des Testens.

Darstellung der Stichproben

Zur Gewinnung der Stichprobe wurden zunächst mehrere Volksschulen in Wien angeschrieben. Im Anschreiben wurden die DirektorInnen über die geplante Durchführung der Untersuchung an den Schulen, die vorgesehene Auswertung und den praktischen Zweck sowie die Ziele der Untersuchungen informiert. Von 70 angeschriebenen Schulen gaben sechs an, Interesse an der Studie zu haben. In drei Fällen stimmte das Schulforum gegen die Durchführung der Studie an ihren Schulen und in einem anderen Fall zog die Schule ihr anfängliches Interesse zurück. An den verbleibenden zwei Volksschulen teilten dann die KlassenlehrerInnen Einverständniserklärungen an die Kinder für deren Eltern (siehe Anhang C) aus, die nach zwei Wochen wieder eingesammelt wurden. 110 Eltern erklärten sich schließlich damit einverstanden, ihre Kinder an der Studie teilnehmen zu lassen. Davon wurden vier aufgrund mangelnden Sprachverständnisses und ein Kind aufgrund einer Rechenschwäche nachträglich ausgeschlossen. Zwei Kinder hatten zum Zeitpunkt der Untersuchung die Schule gewechselt und weitere zwei Kinder fielen aufgrund längerer Krankheit aus der Studie heraus.

Operationalisierung und Messinstrumente

In der vorliegenden Untersuchung wurden sowohl Zusammenhangs- als auch Unterschiedshypothesen untersucht. Die Messinstrumente der beiden Variablen,

tatsächliche Leistung beim adaptiven oder konventionellen Testen und subjektive Einschätzung der eigenen Leistung beim adaptiven oder konventionellen Testen, werden im Folgenden genauer dargestellt.

Adaptives Intelligenz Diagnostikum 3. Die tatsächliche Leistung beim adaptiven und konventionellen Testen wurde in dieser Untersuchung durch Berechnung der Personenparameter von drei Untertests des Adaptiven Intelligenz Diagnostikums 3 (AID 3; Kubinger & Holocher-Ertl, 2014) erfasst. Beim AID 3 handelt es sich um eine Intelligenz-Testbatterie für Kinder im Alter von sechs bis 15 Jahren bei der die meisten der zwölf Untertests standardmäßig adaptiv, im Sinne des *branched testing*, vorgegeben werden. Altersabhängig wird den Kindern bei den adaptiven Untertests des AID 3 zunächst eine Gruppe von Items und dann, leistungsabhängig nach einem festen Verzweigungsschema, noch zwei weitere Gruppen von Items vorgegeben. Bei den drei ausgewählten Untertests handelte es sich um den verbal-akustischen Untertest 1 *Allgemeinwissen*, den verbal-akustischen Untertest 3 *Angewandtes Rechnen* und den manuell-visuellen Untertest 10 *Analysieren und Synthetisieren*. Diese Untertests wurden ausgewählt, um möglichst verschiedene Fähigkeiten der Kinder abzudecken. Beim *Alltagswissen* werden die Kinder gebeten verschiedene offene Fragen zum Allgemeinwissen zu beantworten, um damit die Fähigkeit des Kindes zu testen „sich Sachkenntnisse über Inhalte anzueignen, die in der heutigen Gesellschaft alltäglich sind" (Kubinger & Holocher-Ertl, 2014, S.9). Beim Untertest *Angewandtes Rechnen* werden den Kindern Textaufgaben gestellt, „um weitgehend unabhängig von schulischen Rechenfertigkeiten [zu] prüfen, inwieweit die Testperson bei der Problemlösung alltäglicher Aufgabenstellungen durch entsprechende Schlussfolgerungen die passenden Rechenoperationen anzuwenden imstande ist" (Kubinger & Holocher-Ertl, 2014, S.10). Beim Untertest *Analysieren und Synthetisieren* werden die Kinder gebeten Muster mit Würfeln, die unterschiedliche Flächen aufweisen, nachzulegen, um zu testen, ob sie über die Fähigkeit verfügen „komplexe (abstrakte) Gestalten durch eine geeignete Strukturierung reproduzieren zu können" (Kubinger & Holocher-Ertl, S.12). Bezüglich der Messgenauigkeit werden für den Untertest 1 ein minimaler Schätzfehler von 0.55, für den Untertest 3 ein minimaler Schätzfehler von 0.58 und für den Untertest 10 ein minimaler Schätzfehler von 0.89 ($\alpha = .05$, einseitig) angegeben. In Bezug auf die Validität ist „inhaltliche Gültigkeit aufgrund von Experten-Rating" und „diskriminante Konstruktvalidität mit zahlreiche[n] Leistungstests und etliche[n] Persönlichkeitsfragebogen (für den AID) gegeben"

(Kubinger & Holocher-Ertl, 2014, S. 3). Der höchste Korrelationskoeffizient zwischen dem Untertest 10 des Adaptiven Intelligenz Diagnostikums 2 (AID 2; Kubinger, 2009) und dem Untertest *Figurenanalogien N2* aus dem kognitiven Fähigkeitstest für 4. bis 12. Klasse, Revision, (KFT 4-12+R; Heller & Perleth, 2000) betrug $r = .40$. Außerdem konnte Neumann (2010) zeigen, dass drei Untertests des Wiener Entwicklungstests (WET; Kastner-Koller & Deimann, 2002) Ergebnisse mehrerer Untertests des AID 2, u.a. die Untertests 1 und 3, vorhersagen konnten.

Um die Auswirkungen der standardmäßigen adaptiven Form des AID 3 mit einer konventionellen Form vergleichen zu können, musste mit Hilfe der Schwierigkeitsparameter des AID 3 zunächst eine konventionelle Form zusammengestellt werden. Dabei wurde darauf geachtet die Items so auszuwählen, dass jede Schwierigkeitsstufe des Itempools vertreten war. Es wurden dazu für die Untertests 1 und 3 jeweils 20 Items pro Altersgruppe aus dem großen Itempool des AID 3 ausgesucht, denen beim adaptiven Testen mit dem AID 3 15 Items gegenüberstanden. Für den Untertest 10 wurden pro Altersgruppe zwölf Items aus dem Itempool ausgewählt. Beim adaptiven Testen werden bei diesem Untertest acht Items standardmäßig vorgegeben. Bei beiden Testvarianten wurden die ersten beiden Items als *warming-up* Aufgaben genutzt und nicht in die Beurteilung der Fähigkeiten miteinbezogen. Die Altersgruppen wurden für das konventionelle Testen von der standardmäßig vorgegebenen adaptiven Form des AID 3 übernommen. Es wurden deshalb für die Untertests 1 und 3 unterschiedliche konventionelle Tests mit je 20 Items für die Kinder im Alter von sechs bis sieben, acht bis neun und zehn bis elf zusammengestellt sowie unterschiedliche konventionelle Tests mit je zwölf Items für die Kinder im Alter von sechs bis neun und zehn bis elf für den Untertest 10. Eine etwas höhere Anzahl an Items war beim konventionellen Testen aufgrund der weniger präzisen Personenparameter bei gleicher Itemanzahl, im Vergleich zum adaptiven Testen, nötig. Die ausgewählten Items wurden dann aufsteigend nach ihrer Schwierigkeit vorgegeben, wie allgemein beim konventionellen Testen üblich. Nur beim Untertest 3 wurde einmal bewusst ein schwierigeres Item vor zwei etwas leichteren Items vorgelegt. Der Grund dafür liegt in der Art der Vorgabe der Items bei diesem Untertest. So werden die leichteren Items gewöhnlich mit einer erklärenden Zeichnung vorgeben und die schwierigeren Items ohne Zeichnung. Zwei der ausgewählten Items (Item 11-2 und Item 5-2), die ohne Zeichnung vorgegeben werden, sind aber laut Schwierigkeitsparameter leichter als eines der Items mit einer

erklärenden Zeichnung (Item 3-4). Um den Ablauf beim Testen zu vereinfachen, wurden deshalb zuerst alle Items mit Zeichnung und dann ohne Zeichnung vorgelegt. In der folgenden Tabelle (Tabelle 1) sind die zusammengestellten konventionellen Varianten des AID 3s, die in dieser Untersuchung verwendet wurden, mit ihren Schwierigkeitsparameter in der Reihenfolge, in der sie vorgegeben wurden, dargestellt.

Tabelle 1

Ausgewählte Items für jeweiligen Untertest und Alter der Testpersonen mit Schwierigkeitsparametern

Untertest	Alter			Schwierigkeitsparameter
	6 -7	8-9	10-11	
1	1-1			-9.35
	1-3			-7.87
	1-4			-7.19
		3-1		-6.79
		2-3		-6.12
	2-5	2-5		-5.04
	10-1	10-1		-3.86
	10-3	10-3	10-3	-3.38
	4-1	4-1	4-1	-2.28
	4-2			-2.14
	10-5	10-5	10-5	-1.98
	9-5			-1.87
	3-3			-1.75
	3-5	3-5	3-5	-1.65
	5-1	5-1	5-1	-1.09
	16-3			-0.40
			4-5	-0.36
	10-4	10-4	10-4	-0.03
	11-3	11-3	11-3	0.39
	5-2	5-2	5-2	0.70

	11-4	11-4	11-4	1.90
		12-3	12-3	2.16
		6-3	6-3	2.65
	5-5	5-5	5-5	2.92
		12-4	12-4	3.08
		6-1	6-1	3.22
			7-4	3.66
	5-4	5-4	5-4	4.06
			7-1	4.24
		6-4	6-4	4.69
			7-4	5.17
3	1-1			-7.57
	1-4			-6.58
	2-2	2-2		-6.15
	3-1	3-1	3-1	-5.39
	3-2	3-2	3-2	-4.70
	2-3	2-3		-3.98
	3-3	3-3	3-3	-3.51
	9-4			-2.93
		10-3		-2.73
	2-5	2-5		-2.55
	10-5	10-5	10-5	-2.08
	2-4	2-4		-1.61
	3-5	3-5	3-5	-1.05
	3-4	3-4	3-4	0.54
	11-2	11-2	11-2	-0.64
	5-2	5-2	5-2	0.00
			13-1	0.79
	5-3	5-3	5-3	1.12
	4-4	4-4	4-4	1.45
			11-4	1.74
	5-5	5-5	5-5	2.01
	12-5	12-5	12-5	2.53
	5-4	5-4	5-4	2.95

		12-4	12-4	3.03
			7-3	3.50
		6-4		3.79
			8-2	3.89
			7-5	4.97
			7-4	6.50
10	1-1	1-1		-7.58
	2-1	2-1		-5.35
	1-2	1-2		-4.97
	2-2	2-2	2-2	-4.21
	8-2	8-2	8-2	-3.16
	9-2	9-2	9-2	-1.82
	3-2	3-2	3-2	-1.27
	10-1	10-1		-0.29
	4-1	4-1	4-1	0.02
			5-2	0.76
	10-2	10-2		0.77
			14-1	0.92
	4-2	4-2	4-2	2.15
	5-1	5-1	5-1	2.18
			11-2	2.55
			11-1	2.86
			6-2	3.67

Anmerkung. Die Kennzeichnung der Items entspricht dem Block und der Nummer des jeweiligen Items innerhalb dieses Blockes. Beispielsweise würde das dritte Item des ersten Blocks, in dieser Tabelle, als „1-3“ aufgeführt werden. Die Schwierigkeitsparameter der einzelnen Items des AID 3 wurden nicht veröffentlicht.

Subjektive Einschätzung der eigenen Leistung. Die subjektive Einschätzung der eigenen Leistung (EL) beim adaptiven und konventionellen Testen wurde in beiden Bedingungen und in allen drei Untertests mit drei Fragen anhand einer bipolaren, vierkategorialen, direkten Ratingskala mit aufsteigender Skalenorientierung ohne Skalenmitte gemessen. Die Fragen lauteten:

(1) Wie viele Aufgaben glaubst du, hast du gelöst? Sehr viele, viele, wenige oder sehr wenige? (Einschätzung des Anteils an gelösten Items, EA)
(2) Wie zufrieden bist du mit deiner Leistung im letzten Test? Sehr zufrieden, zufrieden, unzufrieden oder sehr unzufrieden? (Zufriedenheit mit der eigener Leistung, ZL)
(3) Welche Aufgaben würdest du jetzt noch lieber versuchen? Viel schwierigere, schwierigere, leichtere oder viel leichtere? (Motivation der Testperson, MT)

Diese Fragen wurden aus der Sichtung der aktuellsten Studien zu psychischen Reaktionen auf das adaptive und konventionelle Testen gewonnen. Die erste Frage wurde aufgrund der, oben bereits erwähnten, Vermutung, dass Testpersonen ihre Leistung über die Anzahl an gelösten Items unabhängig von deren Schwierigkeit einschätzen, gewählt. In vorangegangenen Studien wurden Testpersonen häufig gebeten den Prozentsatz oder die genaue Anzahl der von ihnen gelösten Items zu schätzen. Da in dieser Studie jedoch Volksschüler befragt wurden, wurde auf eine Schätzung des Prozentsatzes oder der genauen Anzahl verzichtet und stattdessen nur auf einer vierkategorialen Skala nach deren subjektiven Einschätzung gefragt (sehr viele, viele, wenige oder sehr wenige gelöste Aufgaben). In den oben genannten Studien wurde zudem mehrfach eine gemessene positive Korrelation zwischen der Anzahl an gelösten Items und der eigenen Zufriedenheit bzw. der eigenen Motivation gemessen, weshalb die zweite und dritte Frage ebenfalls aufgenommen wurden. Die zweite Frage beruht auf der Annahme, dass Kinder, die ihre Leistungen im Test hoch einschätzen eher mit ihnen Leistungen zufrieden sind als Kinder, die davon überzeugt sind, dass ihre Leistungen eher niedrig einzustufen sind. Die dritte Frage ist zudem stark an die von Kubinger und Holocher-Ertl (2014) selbst formulierte Frage, ob das Kind „nun lieber leichtere, gleichschwierige oder schwierigere Aufgaben versuchen möchte“ (Kubinger & Holocher-Ertl, 2014, S. 37), angelehnt. Dabei wird angenommen, dass ein Kind, das angibt, dass es tendenziell eher schwierige Aufgaben bearbeiten möchte, dies aufgrund einer höheren eigenen Motivation äußert und dass dieses Kind aufgrund besserer Testleistungen motivierter ist. Zur Unterstützung der Verständlichkeit der einzelnen Fragen wurden den Kindern Smileys, passend zu der direkten Ratingskala, vorgelegt.

Durchführung und Ablauf

Die Untersuchungen wurden immer vormittags zwischen acht und zwölf Uhr an den beiden Schulen durchgeführt. In der Volksschule Klausenburger Straße 25 wurde von der Schulleitung ein Raum für die Testung zur Verfügung gestellt. In der Volksschule Pannaschgasse fand die Testung vor den Klassenzimmern statt. In beiden Volksschulen lagen im Allgemeinen keine Lärmbelästigungen oder sonstigen Ablenkungen vor, so dass sich die Kinder ungestört auf die Testung konzentrieren konnten. Dabei wurden immer zwei SchülerInnen gleichzeitig aus der Klasse geholt oder von den KlassenlehrerInnen zur Testung geschickt und parallel, von zwei Testleiterinnen, getestet. Zu Beginn der Testung wurde zunächst der Untertest 12 aus dem AID 3 adaptiv vorgegeben, der Teil einer anderen Studie (Sabo, 2018) war. Danach erfolgte die Vorgabe der, für die hier beschriebene Studie, relevanten drei Untertests. Alle Instruktionen wurden sowohl bei der adaptiven als auch bei der konventionellen Testung wortwörtlich aus dem Manual des AID 3 entnommen. Im Anschluss an die jeweiligen Untertests wurde die Befragung des Kindes zur subjektiven Einschätzung der eigenen Leistung des jeweils letzten Untertests durchgeführt. Daraufhin erfolgte die nochmalige Vorgabe des Untertests 12 für die Studie von Sabo (in Vorbereitung). Die Durchführungszeit für diese beiden Studien schwankte, je nach Kind, zwischen 40 und 60 Minuten, wobei die Vorgabe der konventionellen Form in der Regel nur geringfügig länger dauerte. Im Anschluss an die Testung wurde den Kindern für die Teilnahme gedankt und sie erhielten die Möglichkeit, Fragen zu stellen. Als Belohnung erhielt jedes Kind zwei Bonbons und wurde gebeten zurück in seine Klasse zu gehen.

Beschreibung der statistischen Analyse

Die Berechnung der Personenparameter erfolgte beim adaptiven Testen nach dem Manual des AID 3 (Kubinger & Holocher-Ertl, 2014). Die Personenparameter beim konventionellen Testen wurden mithilfe des R-Packages PP berechnet. Die subjektive Einschätzung der eigenen Leistung entsprach dem Mittelwert der vierkategorialen Antworten auf die drei oben genannten Fragen.

Alle nun folgenden statistischen Berechnungen wurden mittels Microsoft Excel und Rstudio durchgeführt.

Zur Analyse der ersten Hypothese wurde der Wilcoxon-Rangsummen-Test, um die subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen zu vergleichen, durchgeführt. Der parameterfreie Test wurde aufgrund der vierkategorialen direkten Ratingskala, die für die Erfassung der subjektiven Einschätzung der eigenen Leistung verwendet wurde, herangezogen.

Die Hypothesen 2a und 2b wurden durch Spearman-Korrelationen, zwischen der tatsächlichen Leistung und der subjektiven Einschätzung der eigenen Leistung beim konventionellen bzw. adaptiven Testen, analysiert. Auch für die Wahl dieses Verfahrens war die vierkategoriale direkte Ratingskala, die zur Erfassung der subjektiven Einschätzung der eigenen Leisten verwendet wurde, ausschlaggebend.

Die dritte Hypothese wurde ebenfalls mit Hilfe des Wilcoxon-Rangsummen-Tests untersucht. Zuvor wurden die Quartile der Leistungen der Testpersonen berechnet und anschließend die einzelnen Leistungen der Kinder den entsprechenden Quartilen zugeordnet. Ebenso wurden die Quartile der subjektiven Einschätzung berechnet und die einzelnen Einschätzungen der Kinder den entsprechenden Quartilen zugeordnet. Anschließend wurde die Differenz zwischen der Einschätzung der eigenen Leistungen und der tatsächlichen eigenen Leistungen berechnet, um ein Maß der verzerrten Einschätzung der eigenen Leistungen zu erhalten. Als leistungsschwache Testpersonen wurden dabei alle Testpersonen angesehen, deren Leistung sich im untersten Quartil befand, als leistungsstark alle Testpersonen, deren Leistung sich im obersten Quartil befand. Schließlich wurde der Wilcoxon-Rangsummen-Test herangezogen, um zu analysieren, ob ein Unterschied der Verzerrung beim adaptiven und konventionellen Testen vorliegt.

Bei den Hypothesen 4a und 4b wurden wiederum Spearman-Korrelationen angewendet, um den Zusammenhang zwischen tatsächlicher und subjektiv eingeschätzter Leistung bei älteren und jüngeren Kindern beim konventionellen bzw. adaptiven Testen zu untersuchen. Zu den älteren Kindern wurden dabei alle Kinder von neun bis elf und zu den jüngeren alle von sechs bis acht Jahren gezählt.

Ergebnisse

Deskriptivstatistische Auswertung

Deskriptive Beschreibung der Stichprobe. Die Stichprobe umfasste *N* = 101 SchülerInnen, die zur Zeit der Untersuchung eine Volksschulen in Wien besuchten, n_1 = 25 die Volksschule Klausenburger Straße 25 und n_2 = 76 die Volksschule Pannaschgasse. Der Anteil an Schülern betrug dabei 52% und der Mittelwert des Alters lag bei 8.67 Jahren, wobei alle SchülerInnen zwischen sechs und elf Jahre alt waren. Über die genaue Verteilung des Geschlechts und des Alters auf die beiden Experimentalgruppen, adaptives und konventionelles Testen, gibt die folgende Tabelle (Tabelle 2) Auskunft. Wie aus der Tabelle ersichtlich ist, unterschied sich der Anteil des Geschlechts oder der Altersbereich innerhalb der einzelnen Gruppen nicht wesentlich voneinander. So lag der Median des Alters in allen Gruppen bei 9. Im Gegensatz zum adaptiven Testen gab es beim konventionellen Testen sechsjährige Mädchen und keine elfjährigen Jungen. Das Verhältnis zwischen Mädchen und Jungen war in der konventionellen Bedingung etwas weniger ausgeglichen als beim adaptiven Testen (adaptiv: 53% Mädchen; konventionell: 42% Mädchen).

Tabelle 2

Deskriptivstatistische Beschreibung der beiden Experimentalgruppen

Form	N_W	N_M	Alter			
			$Median_W$	$Median_W$	$Altersbereich_W$	$Altersbereich_M$
adaptiv	28	25	9	9	7 – 11	7 – 11
konventionell	20	28	9	9	6 – 11	7 – 10

Anmerkung. N_W = Stichprobenumfang des weiblichen Geschlechts, N_M = Stichprobenumfang des männlichen Geschlechts, $Median_W$ = Median des weiblichen Geschlechts, $Median_M$ = Median des männlichen Geschlechts, $Altersbereich_W$ = Altersbereich des weiblichen Geschlechts, $Altersbereich_M$ = Altersbereich des männlichen Geschlechts

Die SchülerInnen stammten insgesamt aus 26 verschiedenen Ländern. Dabei entfielen die größten Anteile auf Österreich (42 Kinder), Serbien (12 Kinder), Syrien (8

Kinder) sowie kleinere Anteile auf die Türkei (7 Kinder), Ägypten (5 Kinder), Bulgarien (3 Kinder), Kroatien (3 Kinder), Albanien (2 Kinder) und Tschechinnen (2 Kinder). Schließlich kam jeweils noch ein Kind aus Bosnien, China, Griechenland, Indien, Italien, Japan, Kosovo, Litauen, Libyen, Makedonien, Palästina, den Philippinen, Polen, Rumänien, Tunesien, Ungarn und Vietnam.

Beschreibung der subjektiven Einschätzung der eigenen Leistung. Die Schülerinnen schätzen sich, wie in der folgenden Tabelle (Tabelle 3) sichtbar, in jedem Untertest und unter beiden Testbedingungen im Durchschnitt überdurchschnittlich ein. Am schlechtesten schätzen sie sich im Untertest 1 unter beiden Bedingungen und im Untertest 10 in der konventionellen Bedingung, mit einem Median von 2.67, ein. In allen anderen Bedingungen schätzen sich die Testpersonen, mit einem Median von 3.00, besser ein. Die kleinsten Werte der subjektiven Einschätzung der eigenen Leistungen liegen zwischen einem Wert von 1.00 und 1.67. Der größte Wert ist in beiden Bedingungen und allen Untertests, mit Ausnahme des Wertes im ersten Untertest in der adaptiven Bedingung (EL_g = 3.67), der Wert 4.00. Die Werte unterscheiden sich somit ebenfalls nur geringfügig voneinander.

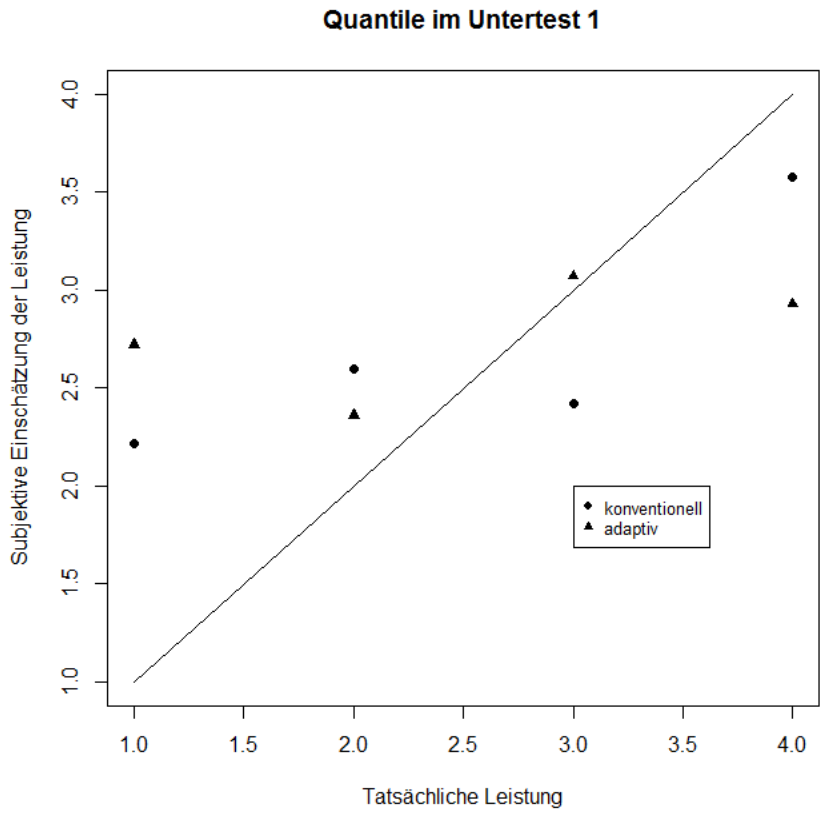
Tabelle 3

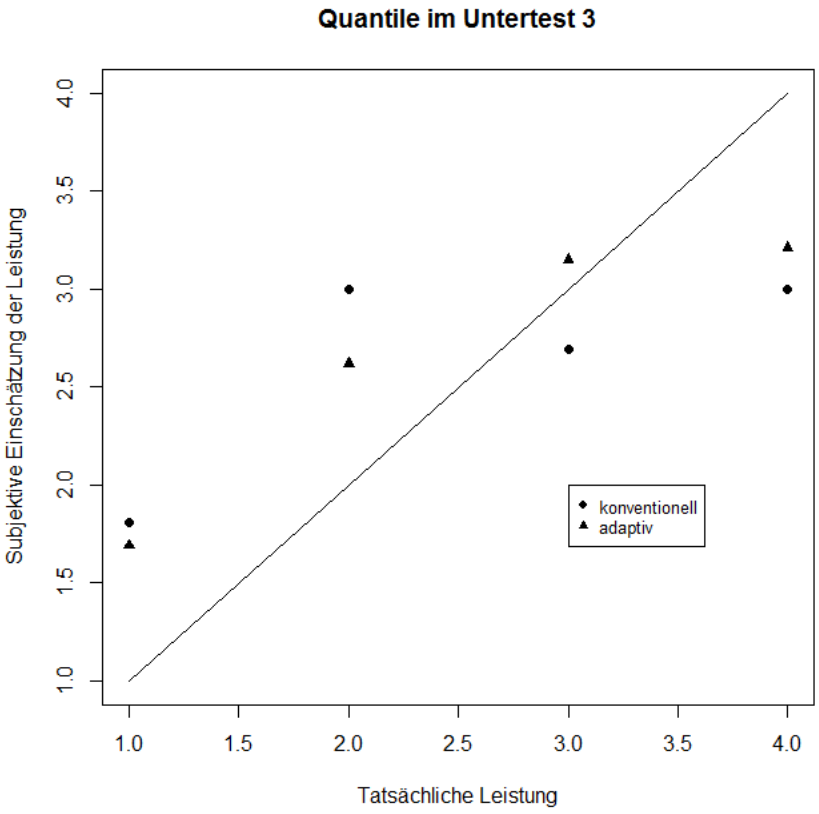
Deskriptivstatische Beschreibung der subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen in den verschiedenen Untertests

Untertest	Form	subjektive Einschätzung der eigenen Leistung		
		Median	EL_k	EL_g
U1	adaptiv	2.67	1.67	3.67
	konventionell	2.67	1.33	4.00
U3	adaptiv	3.00	1.33	4.00
	konventionell	3.00	1.00	4.00
U10	adaptiv	3.00	1.33	4.00
	konventionell	2.67	1.00	4.00

Anmerkung. EL_k = kleinster Wert der subjektiven Einschätzung der eigenen Leistung, EL_g = größter Wert der subjektiven Einschätzung der eigenen Leistung

Die nächste Abbildung (Abb. 1) illustriert die Unterschiede zwischen der Einschätzung der eigenen Leistung und der tatsächlichen Leistung für alle Leistungsquartile beim adaptiven und konventionellen Testen in allen Untertests. Die verzerrte Wahrnehmung der eigenen Leistung scheint sowohl bei leistungsstarken als auch bei -schwachen Testpersonen durchaus erkennbar zu sein. So schienen sich die leistungsschwachen Testpersonen, die Testpersonen deren Leistungen im ersten Quartil liegen, zu überschätzen. Je nach Untertest scheint das Ausmaß der Unterschiede zwischen der subjektiven Einschätzung der eigenen Leistung und der gemessenen Leistung der Testpersonen zu variieren. Am größten wirkt diese Überschätzung im Untertest 1 und 10 in der adaptiven Bedingung. Die Testpersonen deren Leistungen im zweiten und dritten Quartil liegen, scheinen sich, bis auf die in der konventionellen Bedingung im Untertest 3, deren Fähigkeiten dem zweiten Quartil zugeordnet werden können, kaum zu über- oder unterschätzen. Deutlich zu erkennen ist jedoch auch, dass sich die leistungsstarken Testpersonen, die Testpersonen deren Fähigkeiten im vierten Quartil liegen, in allen Untertests, zu unterschätzen scheinen.





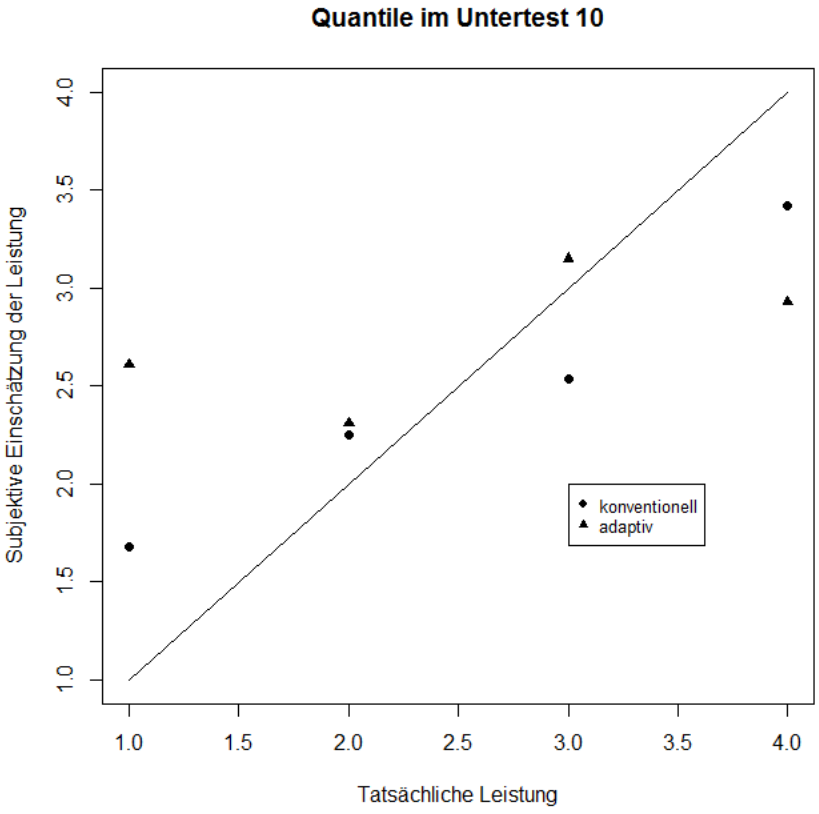


Abbildung 1. Quartile der subjektiven Einschätzung der eigenen Leistung und der tatsächlichen Leistung in den Untertests 1, 3 und 10

Statistische Auswertung der Hypothesen

Hypothese 1. Es können keine signifikanten Unterschiede zwischen der subjektiven Einschätzung der eigenen Leistung (EL) beim adaptiven und konventionellen Testen in den Untertests 1 und 3 gefunden werden und nur ein unbedeutender Unterschied im Untertest 10, der weniger als 9 Prozent der Varianz erklärt (siehe Tabelle 4). Die Hypothese muss deshalb verworfen werden.

Bei genauerer Betrachtung der einzelnen Items der subjektiven Einschätzung der eigenen Leistung (EL) können ebenfalls keine bedeutenden Unterschiede zwischen dem adaptiven und konventionellen Testen beobachtet werden.

Tabelle 4

Unterschiede in der subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen

Untertest	Variabel	$Median_k$	$Median_a$	Ergebnis WR-Test	*p*	*r*	*B*
1	EL	2.67	2.67	*W* = 1272.5	>.999[1]		
	EA	2	2	*W* = 1446.5	.201		
	ZL	3	3	*W* = 1060.0	.112		
	MT	2	2	*W* = 1304.0	.815		
3	EL	3.00	3.00	*W* = 1389.0	.422		
	EA	3	3	*W* = 1346.5	.596		
	ZL	4	3	*W* = 1166.0	.435		
	MT	2	3	*W* = 1548.5	.046	-.20	.04
10	EL	2.67	3.00	*W* = 1701.0	.003	-.29	.08
	EA	3	3	*W* = 1540.5	.057		
	ZL	3	4	*W* = 1480.0	.129		
	MT	2	3	*W* = 1642.0	.009	-.26	.07

Anmerkung. $Median_k$ = Median beim konventionellen Testen, $Median_a$ = Median beim adaptiven Testen, WR-Test = Wilcoxon-Rangsummen-Test, *r* = Korrelationskoeffizient, *B* = Bestimmtheitsmaß (Korrelationskoeffizient und Bestimmtheitsmaß wurden nur bei signifikanten Ergebnissen berechnet); [1]berechneter Wert: *p* = 1.000

Hypothese 2a. Es kann kein signifikant positiver Zusammenhang zwischen der Einschätzung der eigenen Leistung (EL) und der tatsächlichen Leistung im ersten Untertest beim konventionellen Testen gefunden werden (r = .28, p = .054) und lediglich signifikante Zusammenhänge von nicht relevantem Ausmaß im dritten (r = .37, p = .009) und zehnten Untertest (r = .51, p = .0002) beobachtet werden, die weniger als 14 bzw. weniger als 27 Prozent der erklärten Varianz widerspiegeln. Die Hypothese wird somit vorläufig angenommen.

Bei genauerer Betrachtung der einzelnen Items der subjektiven Einschätzung der eigenen Leistung (EL) können ebenfalls keine signifikanten Zusammenhänge von relevantem Ausmaß gefunden werden (s. Tabelle 5). Der größte von ihnen entspricht dabei nur etwa 32 Prozent erklärter Varianz (Untertest 10, EA).

Tabelle 5

Spearman-Korrelation zwischen der tatsächlichen Leistung und der eingeschätzten Leistung beim konventionellen Testen

Untertest	Variabel	*n*	*r*	*B*	*p*
1	EL	48	.28		.054
	EA	48	.39	.15	.007
	ZL	48	-.16		.267
	MT	48	.29	.08	.048
3	EL	48	.37	.14	.009
	EA	48	.36	.13	.011
	ZL	48	.15		.293
	MT	48	.33	.11	.020
10	EL	48	.51	.26	< .001[1]
	EA	48	.57	.32	< .001[2]
	ZL	48	.23		.121
	MT	48	.36	.13	.012

Anmerkung. n = Stichprobengröße, r = Spearman-Korrelationskoeffizient, B = Bestimmtheitsmaß (wurde nur bei signifikanten Ergebnissen berechnet); [1]berechneter Wert: p = .0002, [2]berechneter Wert: p = .000

Hypothese 2b. Es kann keine signifikante Korrelation zwischen der Einschätzung der eigenen Leistung (EL) und der tatsächlichen Leistung im ersten Untertest beim adaptiven Testen gefunden werden (r = .08, p = .587) und nur signifikante Zusammenhänge von nicht relevantem Ausmaß im dritten Untertest (r = .46 p = .00005), der in etwa 21 Prozent der erklärten Varianz entspricht, und im zehnten Untertest (r = .29, p = .036), der weniger als 9 Prozent der Varianz erklärt. Die Hypothese wird somit vorläufig angenommen.

Betrachtet man die einzelnen Items der subjektiven Einschätzung der eigenen Leistung in der adaptiven Bedingung genauer (siehe Tabelle 6), zeigt sich, dass keiner der Zusammenhänge einem bedeutenden Effekt entspricht. So erklärt der größte Effekt nur etwa 18 Prozent der erklärten Varianz (Untertest 10, EA).

Tabelle 6

Spearman-Korrelation zwischen der tatsächlichen Leistung und der eingeschätzten Leistung beim adaptiven Testen

Untertest	Variabel	*n*	*r*	*B*	*p*
1	EL	53	.08		.587
	EA	53	.15		.291
	ZL	53	-.11		.428
	MT	53	.17		.234
3	EL	53	.46	.21	< .001[1]
	EA	53	.40	.16	.003
	ZL	53	.30	.09	.031
	MT	53	.35	.12	.011
10	EL	53	.29	.08	.036
	EA	53	.43	.18	.001
	ZL	53	.24		.090
	MT	53	-.09		.524

Anmerkung. n = Stichprobengröße, r = Spearman-Korrelationskoeffizient, B = Bestimmtheitsmaß (wurde nur bei signifikanten Ergebnissen berechnet); [1]berechneter Wert: p = .00005

Hypothese 3. Es gibt weder signifikante Unterschiede im Ausmaß der Verzerrung zwischen der subjektiven Einschätzung der eigenen Leistung und der tatsächlichen Leistung bei leistungsstarken noch bei leistungsschwachen Testpersonen beim adaptiven und konventionellen Testen (siehe Tabelle 7). Die Hypothese wird somit vorläufig angenommen.

Tabelle 7

Unterschiede in der Verzerrung der subjektiven Einschätzung der eigenen Leistung von leistungsschwachen und leistungsschwachen Personen beim adaptiven und konventionellen Testen

Untertest	Quartil	$Median_K$	$Median_A$	Ergebnis des WR-Test	*p*
1	1.	1.0	2.0	*W* = 63.0	.301
	4.	0.0	-1.0	*W* = 57.0	.134
3	1.	0.0	0.0	*W* = 64.0	.640
	4.	-0.5	-1.0	*W* = 101.5	.605
10	1.	1.0	2.0	*W* = 66.0	.086
	4.	0.0	-1.0	*W* = 58.5	.165

Anmerkung. $Median_K$ = Median der Differenz zwischen tatsächlicher und eingeschätzter Leistung in Quartilen beim konventionellen Testen, $Median_A$ = Median der Differenz zwischen tatsächlicher und eingeschätzter Leistung in Quartilen beim adaptiven Testen, WR-Test = Wilcoxon-Rangsummen-Test

Bei genauerer Betrachtung der Daten zeigt sich jedoch nicht in allen Untertests beim konventionellen (siehe Tabelle 8) und adaptiven Testen (siehe Tabelle 9) ein signifikanter Unterschied zwischen dem Quartil der subjektiven Einschätzung der eigenen Leistungen und dem Quartil der tatsächlichen Leistung bei leistungsschwachen Testpersonen (Testpersonen deren Fähigkeiten im ersten Quartil liegen) bzw. leistungsstarken Testpersonen (Testpersonen deren Fähigkeiten im vierten Quartil liegen).

So gibt es weder bei den leistungsstarken noch bei den leistungsschwachen Testpersonen bedeutende Unterschiede zwischen dem Quartil der subjektiven

Einschätzung der eigenen Leistung und dem Quartil der tatsächlichen Leistung beim konventionellen Testen (siehe Tabelle 8). Nur zwei signifikante Unterschiede können beobachtet werden (Untertest 1, 1. Quartil; Untertest 3, 4. Quartil), wobei beide nur unbedeutenden Effektstärken, die nur in etwa 25 Prozent Varianz widerspiegeln, entsprechen.

Tabelle 8

Unterschiede zwischen der subjektiven Einschätzung der eigenen Leistung und der tatsächlichen Leistung von leistungsschwachen und leistungsschwachen Personen beim konventionellen Testen

Untertest	Quartil	$Median_k$	Ergebnis WVR-Test	*r*	*B*	*p*
1	1.	2.0	*W* = 0	- .50	.25	.034
	4.	4.0	*W* = 10			.089
3	1.	1.0	*W* = 0			.057
	4.	3.0	*W* = 36	- .49	.24	.013
10	1.	2.0	*W* = 0			.095
	4.	4.0	*W* = 10			.098

Anmerkung. M_k = Median der Quartile der eingeschätzten Leistung beim konventionellen Testen, WVR-Test = Wilcoxon-Vorzeichen-Rang-Test, *r* = Korrelationskoeffizient, *B* = Bestimmtheitsmaß (Korrelationskoeffizient und Bestimmtheitsmaß wurden nur bei signifikanten Ergebnissen berechnet)

In der adaptiven Bedingung (siehe Tabelle 9) zeigt sich sowohl im ersten als auch im zehnten Untertest eine signifikante Überschätzung der eigenen Leistung bei den leistungsschwachen Testpersonen. In beiden Fällen handelt es sich aber um unbedeutende Effekte, die im ersten Fall etwa 36 und im zweiten Fall etwa 35 Prozent der Varianz erklären können. Die leistungsstarken Testpersonen in der adaptiven Bedingung unterschätzen sich in jedem Untertest signifikant, aber nur in unbedeutendem Ausmaß. Es handelt sich bei allen Zusammenhängen nur um kleine Effekte, wobei der größte etwa 25 Prozent der Varianz erklärt.

Bedeutende Unterschiede lassen sich im zweiten und dritten Quartil in keinem Untertest und unter keiner der beiden Testbedingungen beobachten. So kann lediglich,

in der konventionellen Bedingung im Untertest 3, ein signifikanter Unterschied zwischen der Leistung der Testpersonen, die im zweiten Quartil liegen, und deren subjektiven Einschätzung der eigenen Leistung (Median = 3) gefunden werden (W = 3, p = .021, r = .49). Dieser unbedeutende Effekt entspricht dabei nur einer erklärten Varianz von etwa 24 Prozent.

Tabelle 9

Unterschiede zwischen der subjektiven Einschätzung der eigenen Leistung und der tatsächlichen Leistung von leistungsschwachen und leistungsstarken Personen beim adaptiven Testen

Untertest	Quartil	$Median_a$	Ergebnis WVR-Test	*r*	*B*	*p*
1	1.	3.0	*W* = 0	-.60	.36	.005
	4.	3.0	*W* = 36	-.47	.22	.013
3	1.	1.0	*W* = 0			.098
	4.	3.5	*W* = 28	-.39	.15	.019
10	1.	3.0	*W* = 0	-.59	.35	.003
	4.	3.0	*W* = 0	-.50	.25	.008

Anmerkung. $Median_a$ = Median der Quartile der eingeschätzten Leistung beim adaptiven Testen, WVR-Test = Wilcoxon-Vorzeichen-Rang-Test, *r* = Korrelationskoeffizient, *B* = Bestimmtheitsmaß (Korrelationskoeffizient und Bestimmtheitsmaß wurden nur bei signifikanten Ergebnissen berechnet)

Hypothese 4a. Es gibt keinen signifikanten Zusammenhang von relevantem Ausmaß zwischen der tatsächlichen Leistung und der subjektiven Einschätzung der eigenen Leistungen (EL) bei jüngeren Kindern (siehe Tabelle 10). Tatsächlich gibt es nur einen signifikanten Zusammenhang im Untertest 3 in der adaptiven Bedingung (r = .45, p = .036), der nur etwa 20 Prozent der Varianz erklärt. Die Hypothese wird somit vorläufig angenommen.

Bei genauerer Betrachtung der einzelnen Items der subjektiven Einschätzung der eigenen Leistung (EL) kann kein signifikanter Zusammenhang von relevantem Ausmaß zwischen ihnen und der tatsächlichen Leistung gefunden werden (siehe Tabelle 10).

Tabelle 10

Spearman-Korrelation zwischen der tatsächlichen Leistung und der eingeschätzten Leistung von jüngeren Kindern beim konventionellen und adaptiven Testen

Untertest	Testform	Variabel	*n*	*r*	*B*	*p*
1	konventionell	EL	19	.04		.855
	adaptiv	EL	22	.17		.461
	konventionell	EA	19	.33		.170
	adaptiv	EA	22	.09		.702
	konventionell	ZL	19	-.15		.549
	adaptiv	ZL	22	.06		.779
	konventionell	MT	19	-.10		.672
	adaptiv	MT	22	.13		.571
3	konventionell	EL	19	-.09		.727
	adaptiv	EL	22	.45	.20	.036
	konventionell	EA	19	.06		.809
	adaptiv	EA	22	.41		.056
	konventionell	ZL	19	-.11		.648
	adaptiv	ZL	22	.46	.21	.030
	konventionell	MT	19	-.10		.694
	adaptiv	MT	22	.22		.329
10	konventionell	EL	19	.44		.059
	adaptiv	EL	22	.12		.585
	konventionell	EA	19	.60	.36	.006
	adaptiv	EA	22	.34		.127
	konventionell	ZL	19	.20		.409
	adaptiv	ZL	22	.20		.379
	konventionell	MT	19	.24		.319
	adaptiv	MT	22	-.19		.396

Anmerkung. *n* = Stichprobengröße, *r* = Spearman-Korrelationskoeffizient, *B* = Bestimmtheitsmaß (wurde nur bei signifikanten Ergebnissen berechnet)

Hypothese 4b. Es gibt keinen signifikant positiven Zusammenhang von relevantem Ausmaß zwischen der tatsächlichen Leistung und der subjektiven Einschätzung der eigenen Leistungen (EL) in der konventionellen und adaptiven Bedingung bei älteren Kindern (siehe Tabelle 11). Der größte Zusammenhang in der konventionellen Bedingung schlägt sich dabei in 30 Prozent der erklärten Varianz nieder (Untertest 10), in der adaptiven Bedingung in nur etwa 22 Prozent der erklärten Varianz (Untertest 3). Die Hypothese wird somit vorläufig angenommen.

Bei genauerer Betrachtung der Zusammenhänge der einzelnen Items der subjektiven Einschätzung der eigenen Leistungen (EL) zeigen sich weder in der konventionellen Bedingung noch in der adaptiven Bedingung signifikante Zusammenhänge von relevantem Ausmaß (siehe Tabelle 11).

Tabelle 11

Spearman-Korrelation zwischen der tatsächlichen Leistung und der eingeschätzten Leistung von älteren Kindern beim konventionellen und adaptiven Testen

Untertest	Testform	Variabel	*n*	*r*	*B*	*p*
1	konventionell	EL	29	.39	.15	.037
	adaptiv	EL	31	-.05		.795
	konventionell	EA	29	.43	.18	.018
	adaptiv	EA	31	.12		.508
	konventionell	ZL	29	-.19		.327
	adaptiv	ZL	31	-.23		.224
	konventionell	MT	29	.54	.29	.002
	adaptiv	MT	31	.17		.362
3	konventionell	EL	29	.52	.27	.004
	adaptiv	EL	31	.47	.22	.008
	konventionell	EA	29	.48	.23	.008
	adaptiv	EA	31	.41	.17	.021
	konventionell	ZL	29	.28		.142
	adaptiv	ZL	31	.10		.369
	konventionell	MT	29	.52	.27	.004
	adaptiv	MT	31	.39	.15	.032

10	konventionell	EL	29	.55	.30	.002
	adaptiv	EL	31	.44	.19	.013
	konventionell	EA	29	.58	.34	.001
	adaptiv	EA	31	.50	.25	.004
	konventionell	ZL	29	.22		.258
	adaptiv	ZL	31	.28		.126
	konventionell	MT	29	.41	.17	.027
	adaptiv	MT	31	-.01		.945

Anmerkung. *n* = Stichprobengröße, *r* = Spearman-Korrelationskoeffizient, *B* = Bestimmtheitsmaß (wurde nur bei signifikanten Ergebnissen berechnet)

Diskussion

In der vorliegenden Studie konnten keine bedeutenden Unterschiede in der subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen festgestellt werden. Es ergeben sich weder beim konventionellen noch beim adaptiven Testen signifikante Zusammenhänge von relevantem Ausmaß zwischen der subjektiven Einschätzung der eigenen Leistung und der tatsächlichen Leistung. Die verzerrte Einschätzung der eigenen Leistungen von leistungsschwachen und -starken Kindern unterscheidet sich in ihrem Ausmaß nicht signifikant zwischen den beiden verschiedenen Testvarianten. Es gibt in beiden Bedingungen keinen signifikanten Zusammenhang von relevantem Ausmaß zwischen der subjektiven Einschätzung der eigenen Leistung und der tatsächlichen Leistung bei jüngeren und älteren Kindern.

Interpretation

Die Frage, ob es einen Unterschied der subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen gibt, kann aufgrund der hier beschriebenen Ergebnisse verneint werden. So können keine bedeutenden Unterschiede zwischen den beiden Bedingungen beobachtet werden. Kinder, die adaptiv getestet werden, schätzen ihre Leistungen somit nicht besser oder schlechter ein als Kinder, die konventionell getestet werden. Bei genauerer Betrachtung der einzelnen Items der subjektiven Einschätzung der eigenen Leistung können ebenfalls keine signifikant bedeutenden Unterschiede zwischen dem adaptiven und konventionellen Testen gefunden werden. So unterschieden sich Kinder, die adaptiv getestet werden, nicht in der Einschätzung des Anteils an von ihnen gelösten Items, ihrer Zufriedenheit mit der eigenen Leistung und ihrer Motivation von Kindern, die konventionell getestet werden. In der aktuellen Forschung wird vor allem die Auswirkung des adaptiven und konventionellen Testens auf die Motivation, wie oben bereits ausführlich dargestellt, immer wieder diskutiert. Während die einen eine höhere Motivation beim adaptiven Testen beobachten konnten (Martin & Lazendic, 2018), fanden andere eine höhere Motivation beim konventionellen Testen (Frey et al., 2009) und wieder andere fanden keinen Unterschied in der Motivation (Ling et al., 2017). Die

Ergebnisse der hier beschriebenen Studie können somit die Ergebnisse von Ling et al. (2017) bestätigen, die ebenfalls keine Unterschiede in der Motivation beim adaptiven und konventionellen Testen finden konnten. Es könnte an dieser Stelle eingewendet werden, dass es in dieser Studie keine bedeutenden Unterschiede zwischen dem adaptiven und konventionellen Testen gibt, weil adaptiv in Form des *branched testing* und nicht des *tailored testing* getestet wurde. Wie oben bereits erwähnt, lassen sich beim *branched testing* über die Anzahl an gelösten Items etwas mehr Information zu der eigenen Leistung gewinnen als beim *tailored testing*. Beim *tailored testing* erhält eine leistungsschwache Testperson gleich nach dem ersten Item ein leichteres, beim *branched testing* geschieht dies erst nach der Vorgabe einer ganzen Gruppe von Items, die für leistungsschwache Testpersonen zu schwierig sein könnten.

Wie erwartet ergeben sich aber weder beim konventionellen noch beim adaptiven Testen signifikante Zusammenhänge von relevantem Ausmaß zwischen der subjektiven Einschätzung der eigenen Leistung und der tatsächlichen Leistung. Ebenso verhält es sich mit den Zusammenhängen zwischen der tatsächlichen Leistung und den einzelnen Items der subjektiven Einschätzung. Diese Ergebnisse widersprechen somit einerseits Ortner et al. (2014), die feststellten, dass die tatsächliche Leistung beim konventionellen Testen ein Prädiktor für die eingeschätzte Leistung ist und bestätigen andererseits die Ergebnisse die Autoren, die ebenfalls beobachten konnten, dass die eingeschätzte Leistung kein Prädiktor für die tatsächliche Leistung in der adaptiven Bedingung war. Die Ergebnisse spiegeln die Ergebnisse von Macan et al. (1994) wider, die beim konventionellen Testen ebenfalls einen signifikanten, aber unbedeutenden, Zusammenhang zwischen den beiden Variablen gefunden haben und der Studie von Powell (1994), in der kein Zusammenhang zwischen der eingeschätzten und tatsächlichen Leistung beim adaptiven Testen gefunden werden konnte.

Diese Ergebnisse sprechen insgesamt dafür, dass es weder beim adaptiven noch beim konventionellen Testen einen bedeutenden Zusammenhang zwischen eingeschätzter und tatsächlicher Leistung gibt. Dass es, wie oben bereits dargestellt, keinen Unterschied in der Einschätzung der beiden Testbedingungen gibt, liegt also nicht daran, dass sich die Kinder in dieser Studie beim adaptiven Testen durch das *branched testing* besser einschätzen konnten als in anderen Studien, in denen im Sinne des *tailored testing* getestet wurden. Vielmehr sind die Kinder weder in der

adaptiven noch in der konventionellen Bedingungen in der Lage, ihre eigenen Leistungen akkurat einzuschätzen.

Es könnte an dieser Stelle argumentiert werden, dass die jüngeren Kinder in der Studie für die schwachen Korrelationen in der konventionellen Bedingung verantwortlich sind, die sich laut Literatur schlechter einschätzen können (Miller, 1987; Stipek, 1981). Aber wie erwartet, können auch dann keine bedeutenden Zusammenhänge zwischen eingeschätzter und tatsächlicher Einschätzung gefunden werden, wenn die Kinder in eine jüngere und eine ältere Altersgruppe unterteilt werden. Demnach gibt es, wie vermutet, auch in der Gruppe der älteren Kinder keine signifikanten Zusammenhänge von relevantem Ausmaß zwischen den beiden Variablen und zwar in beiden Testbedingungen. Diese Ergebnisse entsprechen insgesamt der Literatur, laut der sich ältere Kinder zwar weniger überschätzen als jüngere Kinder, aber ebenfalls keine großen Zusammenhänge zwischen eingeschätzter und tatsächlicher Leistung bei älteren Kindern berichtet werden konnten (Kaderavek, et al., 2004).

Ein Unterschied im Ausmaß der Überschätzung der eigenen Leistung durch leistungsschwache Testpersonen oder der Unterschätzung der eigenen Leistung durch leistungsstarke Testpersonen kann zwischen adaptiven und konventionellen Testen nicht beobachtet werden. Jedoch kann, in dieser Studie, im Widerspruch zu aktueller Literatur (Kim et al., 2016), in der konventionellen Bedingung keine bedeutende Verzerrung der subjektiven Einschätzung der eigenen Leistung bei leistungsschwachen oder leistungsstarken Testpersonen gefunden werden. So konnten in der konventionellen Bedingung keine Effektstärken gemessen werden, die für eine deutliche Verzerrung der subjektiven Einschätzung der eigenen Leistung von leistungsschwachen oder leistungsstarken Testpersonen sprechen, wie sie von Kruger und Dunning (1999) berichtet wurde. In der adaptiven Bedingung zeigen sich auch nur unbedeutende Effekte. Insgesamt scheinen sich leistungsstarke Testpersonen weder beim adaptiven noch beim konventionellen Testen deutlich zu überschätzen und leistungsschwache weder beim adaptiven noch beim konventionellen Testen deutlich zu unterschätzen.

Limitationen

Auch wenn diese Studie über viele Stärken im Vergleich zu anderen Studien verfügt, beispielsweise, dass verschiedene Tests, die unterschiedliche Fähigkeiten der Kinder messen durchgeführt wurden oder dass dieselben Tests in einer adaptiven und konventionellen Form vorgelegt wurden und somit direkt miteinander verglichen werden konnten oder dass die Erfassung der subjektiven Einschätzung nicht nur über die Frage nach der Anzahl an gelösten Items operationalisiert wurde, sollten folgende Limitationen nicht unerwähnt bleiben.

Eine Limitation dieser Studie ist der geringe Stichprobenumfang. Dies betrifft vor allem die Analysen der Hypothesen 3, 4a und 4b, bei denen die Stichprobe nicht nur in adaptiv und konventionell, sondern bei den Hypothesen 4a und 4b auch in Jung und Alt und im Falle der Hypothese 3 in vier Leistungsquartile aufgeteilt wurde. In einer größeren Stichprobe hätte man sich die Unterschiede zwischen den einzelnen Jahrgängen detaillierter anschauen können. Gleiches gilt für die genauere Analyse der Leistungsquartile. In einer größeren Stichprobe hätte man die einzelnen Perzentile genauer untersuchen können. So ist es durchaus denkbar, dass die Kinder im den untersten oder obersten Perzentilen stärker zu einer verzerrten Selbsteinschätzung neigen als die Kinder, deren Leistungen nur im untersten oder obersten Quartil lokalisiert wurden.

Wie oben bereits erwähnt wurden immer zwei Kinder parallel getestet. Selten befanden sich die Kinder dabei in Hörweite voneinander. In den wenigen Fällen in denen dies der Fall war, könnte es sein, dass diese Kinder bei den Fragen zu ihrer subjektiven Einschätzung der eigenen Leistungen angaben, sich selbst besser einzuschätzen als dies tatsächlich der Fall war, um sich vor dem anderen Kind nicht zu blamieren. Jedoch machten die Kinder nicht den Eindruck als würden sie sich Gedanken, um das andere Kind machen und die meisten Kinder zeigten ohnehin nur auf das entsprechende Smiley anstatt eine verbale Antwort zu geben. Dennoch kann nicht ausgeschlossen werden, dass die Antworten mancher Kinder durch die Anwesenheit eines zweiten Kindes beeinflusst wurden.

Ausblick

Neben den oben genannten Limitationen und deren Implikationen, ergeben sich auch folgende Fragestellungen für nachfolgende Studien.

Es stellt sich die Frage, ob und wenn ja, wie sich das *tailored testing* vom *branched testing* in der subjektiven Einschätzung der eigenen Fähigkeiten unterscheidet. Zwar gibt es Studien, die verschieden schwierige Varianten des *tailored testing* miteinander vergleichen, jedoch keine, die die Unterschiede dieser beiden unterschiedlichen Formen des adaptiven Testens auf die subjektive Einschätzung untersucht haben. Eventuell vereint das *branched testing* die Vorteile des konventionellen, beispielsweise die Möglichkeiten vorherige Items derselben Itemgruppe zu verbessern, mit denen des adaptiven Testens, beispielsweise kürzere Testlängen und präzisere Personenparameterschätzungen.

Außerdem wäre es interessant den Einfluss von Selbstbewusstsein und anderen Persönlichkeitsmerkmalen auf die subjektive Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen zu untersuchen. Möglicherweise wirken die beiden Testbedingungen unterschiedlich auf Kinder, je nachdem, ob diese mehr oder weniger selbstbewusst sind.

Zu guter Letzt sollte auch die subjektive Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen von älteren Kindern bzw. Jugendlichen miteinander verglichen werden. Möglicherweise können sich Jugendliche in der konventionellen Bedingung besser einschätzen als Volksschulkinder, wodurch sich dann eventuell Unterschiede zwischen dem adaptiven und konventionellen Testen ergeben würden.

Conclusio

Die Ergebnisse dieser Arbeit sprechen insgesamt dafür, dass es keine bedeutenden Unterschiede in der subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen gibt. Auch wenn an dieser Stelle nicht von weiteren eindeutigen Vorteilen des adaptiven Testens berichtet werden kann, so konnten auch keine Nachteile des adaptiven Testens gegenüber dem konventionellen Testen beobachtet werden. Tatsächlich scheinen Kinder ihre Leistungen weder beim

adaptiven noch beim konventionellen Testen gut einschätzen zu können und Verzerrungen der eigenen Selbsteinschätzung von leistungsstarken und leistungsschwachen Kindern treten beim adaptiven Testen nicht in unterschiedlichem Ausmaß zum konventionellen Testen auf. Sollte es tatsächlich keine gravierenden Unterschiede in den psychischen Reaktionen auf diese beiden Testarten geben, spräche dies jedoch eindeutig für das adaptive Testen, das bei einem deutlich geringeren Zeitaufwand präzisere Schätzungen der Fähigkeiten von Personen ermittelt.

Literaturverzeichnis

Alvarez, V., & Adelman, H. S. (1986). Overstatements of self-evaluation by students with psychoeducational problems. *Journal of Learning Disabilities, 19*, 567–571. doi:10.1177/002221948601900910

Bem, D. J., & Lord, C. G. (1979). Template matching: A proposal for probing the ecological validity of experimental settings in social psychology. *Journal of Personality and Social Psychology, 37*, 833 – 846. doi:10.1037//0022-3514.37.6.833

Bergstrom, B. A., & Lunz, M. E. (1999). *CAT for certification and licensure*. In F. Drasgows & J. B. Olsens (Hrgs.), Innovations in computerized assessment (S. 67–91). Mahwah: Erlbaum.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education, 5*, 137–149. doi:10.1207/s15324818ame0502_4

Brown, J. D. (2007). *The self*. New York: Psychology Press.

Brown, J. D., & Gallagher, F. M. (1992). Coming to terms with failure: Private self-enhancement and public self-effacement. *Journal of Experimental Social Psychology, 28*, 3 – 22. doi:10.1016/0022-1031(92)90029-j

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationship between race, test performance, face validiy perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*, 300 – 310. doi:10.1037//0021-9010.82.2.300

Chen, S. Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24*, 241 – 255. doi:10.1177/01466210022031705

Dunning, D., Heath, C., Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*, 69-106. doi:10.1111/j.1529-1006.2004.00018.x

Eggen, T. J. H. M. & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30*, 379–393. doi:10.1177/0146621606288890

Everson, H. T., & Tobias, S. (1998). The ability to estimate knowledge and performance in college: A metacognitive analysis. *Instructional Science, 26*, 65 – 79. doi:10.1007/978-94-017-2243-8_4

Fagot, B. L., & O'Brien, M. (1994). Activity level in young children: Cross-age stability, situational influences, correlates with temperament, and the perception of problem behaviors. *Merrill Palmer Quarterly, 40*, 378 – 398. Retrieved from http://www-jstor-org.uaccess.univie.ac.at/stable/23087351

Fayers, P.M. (2007). Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. *Quality of Research, 16*, 187 – 194. doi:10.1007/s11136-007-9197-1

Felson, R. B. (1981). Ambiguity and bias in the self-concept. *Social Psychology Quarterly, 44*, 64 – 69. doi:10.2307/3033866

Frey, A., & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Messung von Bildungsstandards in Mathematik. *Zeitschrift für Erziehungswissenschaften, Sonderheft 8*, 169 – 184. doi:10.1007/978-3-531-90865-6_10

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. *Diagnostica, 55*, 20 – 28. doi:10.1026/0012-1924.55.1.20

Gibbons, R.D., Weiss, D.J., Kupfer, D.J., Frank, E., Fagiolini, A. Grochocinski, V. J., … Immekus, J.C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*, 361 – 368. doi:10.1176/appi.ps.59.4.361

Hau, K. T., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Education Measurement, 38*, 249 – 266. doi:10.1111/j.1745-3984.2001.tb01126.x

Heller, K. H., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4-12+ R).* Göttingen: Beltz Test.

Kaderavek, J. N., Gillam, R. B., Ukrainetz, T. A., Justice, L. M., & Eisenberg, S. N. (2004). School-age children's self-assessment of oral narrative production. *Communication Disorders Quarterly, 26*, 37 – 48. doi:10.1177/15257401040260010401

Kastner-Koller, U., & Deimann, P. (2012). *Der Wiener Entwicklungstest. Ein Verfahren zur Erfassung des allgemeinen Entwicklungsstandes bei Kindern von 3 bis 6 Jahren.* Göttingen: Hogrefe.

Kim, Y.-H. , Kwon, H., Lee, J., & Chiu, C.-Y. (2016). Why do people overestimate or underestimate their abilities? A cross-culturally valid model of cognitive and motivational processes in self-assessment biases. *Journal of Cross-Cultural Psychology, 47*, 1201 – 1216. doi:10.1177/0022022116661243

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121 – 1134. doi:10.1037/0022-3514.77.6.1121

Kubinger, K. D. (2003). *Adaptives Testen.* In K. D. Kubinger & R. S. Jäger (Hrsg.), Schlüsselbegriffe der Psychologischen Diagnostik (S. 1- 9). Weinheim: Beltz.

Kubinger, K. D. (2009). *Adaptives Intelligenz Diagnostikum - Version 2.2 (AID 2) samt AID 2-Türkisch*. Göttingen: Beltz.

Kubinger, K. D., & Holocher-Ertl, S. (2014). *Adaptives Intelligenz Diagnostikum 3 (AID 3)*. Weinheim: Beltz.

Kubinger, K. D., & Wild, B. (1989). *Die Optimierung der Messgenauigkeit beim "branched" – adaptiven Testen*. In K. D. Kubinger (Hrsg.), Moderne Testtheorie – Ein Abriss samt neusten Beiträgen (2. Aufl., S. 187- 218). Weinheim: Beltz.

Larwood, L., & Whittaker, W. (1977). Managerial myopia: Self-serving biases in organizational planning. *Journal of Applied Psychology, 62*, 194 – 198. doi:10.1037//0021-9010.62.2.194

Legg, S. M., & Buhr, D.C. (2005). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice*, *11*, 23-27. doi:10.1111/j.1745-3992.1992.tb00237.x

Licht, B.G. (1992). *The achievement-related perceptions of children with learning problems: A developmental analysis.* In D. H. Schunk & J. L. Meece (Hrsg.), Student perceptions in the classroom (pp. 247–264). Hillsdale: Erlbaum.

Ling, G., Attali, Y., Finn, B., Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement, 41*, 495 – 511. doi:10.1177/016621617707556

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.

Macan, T.H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology, 47*, 715 – 738. doi: 10.1111/j.1744-6570.1994.tb01573.x

Martin, A. J. (2007). Examining a multidimensional model of student motivation and engagement using a construct validation approach. *British Journal of Educational Psychology, 77*, 413–440. doi:10.1348/000709906X118036

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *American Psychological Association, 110*, 27 – 45. doi:10.1037/edu0000205

Meijer, R. R., & Nering, M. L. (1999). Computerized Adaptive Testing: Overview and Introduction. *Applied Psychological Measurement, 23*, 187 – 194. doi:10.1177/01466219922031310

Miller, A. (1987). Changes in academic self-concept in early school years: The role of conceptions of ability. *Journal of Social Behavior and Personality, 2*, 551–558.

Moe, K. C., & Johnson, M. F. (1988). Participants' reactions to computerized testing. *Journal of Educational Computing Research, 4*, 79 – 86. doi:10.2190/qbqg-mrpx-hfd8-uvrp

Moreland, R., Miller, J., & Laucka, F. (1981). Academic achievement and self-evaluations of academic performance. *Journal of Educational Psychology, 73*, 335 – 344. doi:doi.org/10.1037//0022-0663.73.3.335

Moreno, K. E. Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding armed services vocational aptitudes battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 23*, 187 – 194. doi:10.1177/014662168400800203

Neumann, G. (2010). *Kann aus dem WET für 5-Jährige das Ergebnis im AID 2 als 6-Jähriger prognostiziert werden?* (Diplomarbeit).

Nicholls, J. G. (1978). The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. *Child Development, 49*, 800–814. doi:10.2307/1128250

Nicholls, J. G. (1979). Development of perception of own attainment and causal attributions for success and failure in reading. *Journal of Educational Psychology, 71*, 94–99. doi:10.1037//0022-0663.71.1.94

Ortner, T. M., & Caspers, J. (2011). Consequences of Test Anxiety on Adaptive Versus Fixed Item Testing. *European Journal of Psychological Assessment 27*, 157 – 163. doi:10.1027/1015-5759/a000062

Ortner, T. M.,Weißkopf, E., & Gerstenberg, F. R. (2013). Skilled but unaware of it: CAT undermines a test taker's metacognitive competence. *European Journal of Psychology of Education, 28*, 1–15. doi:10.1007/s10212-011-0100-7

Ortner, T. M., Weißkopf, E., & Koch, T. (2014). I will probably fail. Higher ability students' motivational experiences during adaptive achievement testing. *European Journal of Psychological Assessment, 30*, 48 – 56. doi:10.1027/1015-5759/a000168

Overton, R. C., Harms, H. J., Taylor, L. R., & Zickar, M. J. (1997). Adapting to adaptive testing. *Personnel Psychology, 50*, 171 – 185. doi:10.1111/j.1744-6570.1997.tb00907.x

Powell, Z. E. (1994). The psychological impacts of computerized adaptive testing methods. *Educational Technology, 34*, 41 – 47. Retrieved from http://www-jstor-org.uaccess.univie.ac.at/stable/44428229

Schmidt, F. L., Urry, V. W., & Gugel, J. F. (1978). Computer assisted tailored testing: Examinee reactions and evaluations. *Educational and Psychological Measurement, 38*, 265 – 273. doi:10.1177/001316447803800208

Shaughnessy, J. J. (1979). Confidence judgement accuracy as a predicator of test performance. *Journal of Research in Personality, 13*, 505 – 514. doi:10.1016/0092-6566(79)90012-6

Sabo, T. (in Vorbereitung). *Umsetzung des "uralt" Lerntestkonzepts von Guthke beim AID 3* (Masterarbeit).

Sinkavich, F. J. (1995). Performance and metamemory: Do students know what they don't know? *Journal of Instructional Psychology, 22*, 77 – 87.

Smither, J. W., Millsap, R. E., Reilly, R. R., & Pearlman, K. (1996). An experimental test of the influence of selection procedures on fairness perception attitudes about the organization and job pursuit intentions. *Journal of Business and Psychology, 10*, 297 – 318. doi:10.1007/bf02249605

Stipek, D. J. (1981). Children's perceptions of their own and their classmates' ability. *Journal of Educational Psychology, 73*, 404 – 410. doi:10.1037//0022-0663.73.3.404

Stone, C. A. (1997). Correspondences among parent, teacher, and student perceptions of adolescents' learning disabilities. *Journal of Learning Disabilities, 30*, 660 – 669. doi:10.1177/002221949703000610

Stone, C. A., & May, A. L. (2002). The accuracy of academic self-evaluations in adolescents with learning disabilities. *Journal of Learning Disabilities, 35*, 370 – 383. doi:10.1177/00222194020350040801

Thissen, D., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. In H. Wainer (Hrsg.), Testing algorithms (S. 101 – 133). Hillsdale: Erlbaum.

Tonidandel, S., & Quinones, M. A. (2000). Psychological reactions to adaptive testing. *International Journal of Selection and Assessment, 8*, 7 – 15. doi:10.1111/1468-2389.00126

Tonidandel, S., Quinones, M. A. & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87*, 320 – 332. doi:10.1037//0021-9010.87.2.320

Van der Linden, W. J., & Glass, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. St. Paul: Assessment Systems Corporation.

Veerkamp, W. J. J., & Berger, M. P. F.(1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203 – 226. doi:10.2307/1165378

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education, 7*, 53 – 79. doi:10.1207/s15324818ame0701_5

Wainer, H. (2000). *Introduction and history.* In H. Wainer (Hrsg.), Computerized adaptive testing: A primer (S. 1 – 21). Hillsdale: Erlbaum.

Wainer, H., & Eignor, D. (2000). *Caveats, pitfalls, and unexpected consequences of implementing large scale computerized testing*. In H. Wainer (Hrsg.), Computerized adaptive testing: A primer (S. 271 – 299). Hillsdale: Erlbaum.

Wang, T. Y., & Vispoel, W. P.(1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement 35*, 109 – 135. doi:10.1111/j.1745-3984.1998.tb00530.x

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied *Psychological Measurement, 6*, 473 – 492. doi:10.1177/014662168200600408

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counselling and education. *Measurement and Evaluation in Counseling and Development, 37*, 70 – 84. doi:10.1080/07481756.2004.11909751

Weiss, D. J., & Kingsburgy, G.G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361 – 375. doi:10.1111/j.1745-3984.1984.tb01040.x

Zickar, M.J., Overton, R. C., Taylor, L. R., & Harms, H. J. (1999). *The development of a computerized adaptive selection system for computer programmers in a financial services company*. In F. Drasgow & J. B. Olsen (Hrsg.), Innovations in computerized assessment (S. 7 – 33). Mahwah: Erlbaum.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang A

Abstract (deutsche Fassung)

Die vorliegende Studie untersuchte, ob es Unterschiede in der subjektiven Einschätzung der eigenen Leistung beim adaptiven und konventionellen Testen und ob es Zusammenhänge zwischen der eingeschätzten und der tatsächlichen Leistung gibt. Außerdem wurde untersucht, ob ältere Kinder, im Gegensatz zu jüngeren Kindern, ihre Leistungen akkurat einschätzten können und ob es Unterschiede im Ausmaß des Dunning-Kruger-Effekts beim adaptiven und konventionellen Testen gibt.

Mit Hilfe seines selbsterstellten Fragebogens zur subjektiven Einschätzung der eigenen Leistung und drei Untertests des AID 3, die entweder konventionell oder adaptiv an zwei Wiener Volksschulen vorgegeben wurden, wurden Wilcoxon-Rangsummen-Tests und Spearman-Korrelationen berechnet.

Es findet sich kein bedeutender Unterschied der subjektiven Einschätzung sowie kein signifikanter Unterschied im Ausmaß des Dunning-Kruger-Effekt beim adaptiven und konventionellen Testen. Es kann kein signifikanter Zusammenhang von relevantem Ausmaß zwischen der eingeschätzten und tatsächlichen Leistung beim adaptiven und konventionellen Testen beobachtet werden. Außerdem können sich ältere und jüngere Schüler weder beim adaptiven noch beim konventionellen Testen akkurat einschätzen.

Diese Ergebnisse sprechen dafür, dass die Unterschiede einer adaptiven und konventionellen Testung keinen großen Einfluss auf die subjektive Einschätzung der eigenen Leistung von Kindern haben.

Anhang B

Abstract (englische Fassung)

The present study investigated if there are differences in the subjective estimation of one's own test performance between adaptive and conventional testing and if there are relationships between estimated and actual performance using adaptive and conventional testing. Furthermore it was examined if older children are able to estimate their own test performance precisely in contrast to younger ones and if there are differences in the extent of the dunning-kruger effect between adaptive and conventional testing.

Primary school children were presented with either adaptive or conventional subtests of the AID 3 and asked questions about their estimated test performances. Wilcoxon-rang-sum-tests and Spearman correlations were conducted.

There is no important difference in the subjective estimation of one's own test performance and no differences in the extent of the dunning-kruger effect between adaptive and conventional testing. There is no strong relationship between estimated and actual performance using adaptive testing or conventional testing. Younger and older children were not able to estimate their test performances accurately in the conventional and the adaptive condition.

Based on these results it seems the subjective estimation of children's own test performance is not heavily influenced by the differences of adaptive and conventional testing.

Anhang C

Einverständniserklärung der Eltern

Liebe Eltern und Erziehungsberechtigte,

viele Kinder werden während ihrer Schullaufbahn aus den unterschiedlichsten Gründen mit psychologischen Verfahren untersucht. Damit die Schullaufbahnberatung und Schulpsychologie immer auf dem neuesten Stand der Wissenschaft ist und Verbesserungen der psychologischen Tests vorgenommen werden können, ist die Forschung in diesem Bereich sehr wichtig.
Im Rahmen unserer Masterarbeiten an der Fakultät für Psychologie an der Universität Wien führen wir beide, unter der Leitung von Univ.-Prof. i.R. Dr. Mag. Klaus D. Kubinger, zwei Studien durch, in deren Rahmen wir in den nächsten Wochen mit Einverständnis der Direktion, in der Klasse Ihres Kindes eine Testung mithilfe des AID 3 (Adaptives Intelligenz Diagnostikum) durchführen werden. Speziell für diese Testung werden den Kindern: Textrechenaufgaben vorgegeben, Figurenfolgen vorgegeben, die man um eine weitere Figur ergänzen soll, Würfel zur Verfügung gestellt, bei denen man geometrische Muster anhand einer Vorlage nachbauen soll und Fragen zum Alltagswissen gestellt.

Bei der ersten Studie geht es darum, wie Kinder je nach Art der Testung ihre eigene Leistung beim Testen einschätzen. Dabei werden zwei Arten der Testung miteinander verglichen: das adaptive Testen, bei dem die Aufgabenauswahl während der Testung an die Fähigkeit des Kindes angepasst wird und das konventionelle Testen, bei dem alle Kinder die gleichen Aufgaben bekommen. Interessant ist der Unterschied in der Selbsteinschätzung deshalb, weil sie u.a. Einfluss auf die Motivation der Kinder während der Testung und damit auch auf die gemessene Leistung selbst haben könnte.

Für die zweite Studie wurde für eine der vorher aufgeführten Aufgabengruppen ein spezieller Lerntest entwickelt. Ein Lerntest soll die Lernfähigkeit und das zukünftige Potential des Kindes ermitteln. Das Ziel dieser Studie ist es zu prüfen, ob mit dem entwickelten Lerntest auch wirklich die Lernfähigkeit erfasst werden kann.

Wir wenden uns daher mit der Bitte an Sie, Ihr Kind an dieser Untersuchung teilnehmen zu lassen, vorausgesetzt natürlich, dass Sie und Ihr Kind damit einverstanden sind. Die Untersuchung findet während der Schulzeit statt und dauert zwischen einer halben und einer Stunde. Aus Erfahrung macht den Kindern die Mitarbeit an den Aufgaben viel Spaß. Die gewonnenen Daten werden im Sinne des Datenschutzrechtes ausschließlich für wissenschaftliche Zwecke genutzt. Die Anonymisierung sämtlicher Ergebnisse der SchülerInnen findet bereits während der Testung statt. Wir bitten Sie, bis _____, mit Ihrer Unterschrift auf dem Abschnitt unten Ihr Einverständnis zur Teilnahme Ihres Kindes an den beschriebenen Untersuchungen zu erteilen.

Mit freundlichen Grüßen und herzlichem Dank im Voraus!

XXX & Theresa Sabo

----------bitte abtrennen----------

Ich erkläre mich an der Teilnahme meiner Tochter/ meines Sohnes _____, die oder der die Klasse ____ besucht, geboren am _____, an der Testung der Masterarbeiten von XXX & Theresa Sabo einverstanden.

Datum, Unterschrift _____