Bewährungskontrolle des A2-Verfahrens: Hintergründe, Ergebnisse und Konsequenzen

Benjamin Haarhaus/Dr. Stephan Buchhester/Nina Ristel

Zusammenfassung

Die Testergebnisse von 51 Bewerberinnen und Bewerbern, die in den Jahren 2006 und 2007 das schriftliche A2-Verfahren der DGP durchlaufen haben, werden verwendet, um den Ausbildungs- und Studienerfolg im gehobenen Verwaltungsdienst und vergleichbaren Laufbahnen vorherzusagen. Unkorrigierte Validitätskoeffizienten zwischen r = .42 und r = .52 verdeutlichen, dass der A2 in der Lage ist, den Ausbildungs- und Studienerfolg über mehrere Jahre valide vorherzusagen.

1. Wieso Bewährungskontrollen unabdingbar sind

„Was du nicht messen kannst, kannst du nicht lenken.“ (Robert Kaplan)
Nennen Sie es steuern, beeinflussen, regeln, controllen oder einfach nur mit dem, was Sie tun, exakt sein – die Messbarkeit entscheidet darüber, ob das eigene Handeln zielführend, erfolgreich und beeinflussbar ist. Das gilt für gut quantifizierbare Ergebnisse im Sport, wie z. B. Weite, Dauer und Höhe, genauso wie für schwerer erfassbare Phänomene – wie zum Beispiel der Vorhersage des beruflichen Erfolgs auf der Basis eines Testverfahrens. In der Eignungsdiagnostik stellt die Genauigkeit der Vorhersageergebnisse eine erhebliche und sich ständig ändernde Herausforderung dar.
Berufliche Grund-Folge-Beziehungen werden immer multikausaler und somit immer komplexer. Erinnern wir uns für einen Moment zurück: Im Märchen bekam der älteste Sohn die Mühle. Weil er dem Vater am längsten bei der Bewirtschaftung der Mühle geholfen hatte, hatte der Älteste die meiste Erfahrung. Eignungsdiagnostisch gesehen eine Arbeitsprobe in Form eines Langzeitpraktikums, eines der diagnostischen Instrumente mit der höchsten Vorhersagegenauigkeit (Schmidt & Hunter, 1998).
Wie gestaltet sich die Situation in der heutigen Zeit? Die Mühle ist längst ein Unternehmen mit vielen Produkten an unterschiedlichen Standorten und mit einem weltweit agierenden Einkauf der Rohstoffe oder eine öffentliche Verwaltung mit vielfältigem Dienstleistungsangebot. Jetzt reicht es plötzlich nicht mehr aus, dass der älteste Sohn des Eigentümers am Familienstammsitz die Geschicke der Firma übernimmt. Es werden viel komplexere Anforderungen gestellt. An welchen Standorten sind ggf. ebenfalls „älteste Söhne“ (oder Töchter), an welchen Standorten hat sich wer spezialisiert? Und welches Instrument ist geeignet, diese komplexen Anforderungen so zu messen, dass die Person mit der höchsten Ausprägung im Ergebnis dann auch tatsächlich „das Beste“ aus der Firma macht? Und was ist „das Beste“ überhaupt? Fragen, mit denen sich Eignungsdiagnostiker täglich aktiv auseinandersetzen müssen.
Der Schwierigkeitsgrad der Messbarkeit, die Komplexität der Methoden sowie der Aufwand der Datenerfassung steigen, je lebensnäher und somit komplexer die Zusammenhänge werden und je weniger fassbar die Einflussgrößen sind. Das ist insbesondere bei personalwirtschaftlichen Fragestellungen von besonderer Bedeutung. Hier obliegt dem Anwender der Eignungsdiagnostik im besonderen Maße eine Fürsorgepflicht für die Bewerberinnen und Bewerber.

Die Literatur zu dem Thema ist vielschichtig, vielfältig und ausgiebig diskutiert. Einigkeit herrscht vor allem hinsichtlich der Feststellung, dass noch immer eine erhebliche Kluft zwischen den Anforderungen der Wissenschaft an ein eignungsdiagnostisches Instrument und der betrieblichen Machbarkeit besteht. Praktische Probleme der Datensammlung aus Personalakten, der Gruppenvergleichbarkeit und/oder Stabilität von Einflussfaktoren auf z. B. die Auszubildenden gestalten eine Vorhersage über den Zusammenhang zwischen einem Eignungstest vor dem Ausbildungsstart und dem erfolgreichen Ausbildungsabschluss schwierig, völlig unabhängig davon, welche Parameter als Erfolg gesehen werden. So mag es z. B. für manche Organisationen von Vorteil sein, wenn Auszubildende nur mittelmäßige Prüfungsleistung erbringen, da diese mit größerer Wahrscheinlichkeit in der Organisation bleiben, als jemand mit herausragenden Ergebnissen und einer hohen Fluktuationsneigung. Das bedeutet, dass das erforderliche Außenkriterium als Erfolgsparameter gesellschaftlichen und wirtschaftlichen Veränderungen unterworfen sein kann und sich damit unmittelbar und ggf. ergebnisbeeinträchtigend auf die innerbetriebliche Sozialisation der Auszubildenden niederschlägt.
Kurzum: Es gilt Instrumente zu entwickeln, die aufgrund einer im Vorfeld zu erfolgenden Testung, den Erfolg des getesteten Merkmals besonders gut voraussagen. Dies gestaltet sich mit Zunahme der Zeit zwischen Testzeitpunkt und Erfolgseintritt, mit der Menge der Einflussfaktoren und mit der eingeschränkten Greifbarkeit des zu messenden Parameters immer schwieriger. Die Überprüfung, wie gut das jeweilige Instrument diesem Anspruch gerecht wird, nennt sich Validitäts- oder Bewährungskontrolle.
Seit Jahrzehnten stellt sich die DGP satzungsgemäß dieser gesamtgesellschaftlichen Herausforderung und investiert als eines der wenigen Unternehmen am Markt erhebliche personelle und finanzielle Mittel in diese kontinuierliche Verbesserung der Auswahlinstrumente. Dabei orientiert sich die DGP mit Ihren Verfahren immer an den Standards, die für diesen Bereich definiert wurden: der DIN 33430.
Der Großteil aller Dienstleistungen im Bereich der Personalauswahl wird von Anbietern erbracht, die weder für die DIN 33430 zertifiziert sind, noch im Rahmen einer Selbstverpflichtung (wie z. B. der DGP-Satzung) die harten Kriterien der qualitativ gesicherten Eignungsdiagnostik als internen Standard definieren. Bedenkt man, dass Durchführung und Interpretation von Bewährungskontrollen eine schwierige und erklärungsbedürftige und somit eine oftmals vom Markt nicht refinanzierte Dienstleistung darstellen, sollte dies nicht weiter verwundern.

2. Eine Bewährungskontrolle des DGP A2-Verfahrens

2.1. Hintergrund

Das A2-Verfahren wird zur Bewerberauswahl im gehobenen Verwaltungsdienst und vergleichbaren Laufbahnen eingesetzt. Geprüft werden numerische und verbale Verarbeitungskapazität, Arbeitseffizienz, Rechtschreibung sowie verschiedene Kenntnisbereiche. Die verschiedenen Leistungsbereiche werden laufbahnspezifisch gewichtet und zu einem Gesamtpunktwert verrechnet. Aus dem Gesamtpunktwert wird ein Empfehlungsgrad abgeleitet, der die Eignung der Bewerberinnen und Bewerber für die spezifische Laufbahn auf einer Skala von 1 bis 5 angibt. Ein Empfehlungsgrad von 1 bedeutet, dass die Leistung des Bewerbers oder der Bewerberin den Anforderungen nicht genügt; bei einem Empfehlungsgrad von 5 werden die Anforderungen erfüllt. Der

Empfehlungsgrad prognostiziert somit den Ausbildungs- und Studienerfolg. Die Prognosegüte des Empfehlungsgrades lässt sich statistisch durch eine Korrelation ermitteln, die in diesem Falle auch als Validitätskoeffizient bezeichnet wird. Je höher der Validitätskoeffizient, umso besser sagt der Empfehlungsgrad den Ausbildungserfolg vorher.

2.2. Vorgehen

Die Datenerhebung fand von Mai bis September 2011 statt. Per E-Mail baten wir unsere Kunden, uns (in anonymisierter Form) die Abschlussnoten der Bewerberinnen und Bewerber zukommen zu lassen, die in den Jahren 2006 und 2007 das schriftliche A2-Verfahren durchlaufen und ihre Ausbildung bzw. ihr Studium abgeschlossen haben. Da die Abschlussnoten auf verschiedenen Skalen (Schulnoten von 1 bis 5 und Punktwerte von 0 bis 15) vorlagen, wurden diese aus Gründen der Anschaulichkeit zunächst ineinander überführt, so dass jede/r Bewerber/in bzw. Auszubildende einen Notenwert und einen Punktwert besitzt. Um die Aussagefähigkeit der Ergebnisse sicherzustellen, wurden die Daten anschließend auf Vollständigkeit und Vergleichbarkeit der eingesetzten Verfahren hin untersucht. Nicht verwertbare Datensätze wurden aus der Stichprobe entfernt. Prüfungswiederholer sowie Bewerberinnen und Bewerber, die infolge ärztlich attestierter gesundheitlicher Probleme durch die Abschlussprüfung gefallen sind, wurden nicht berücksichtigt. Zur Analyse der Daten wurden Kreuztabellen und bivariate Korrelationskoeffizienten verwendet. Als zweiter, alternativer Prädiktor dient neben dem Empfehlungsgrad auch der Gesamtpunktwert im A2-Verfahrens. Ein zweites Außenkriterium stellt das Ergebnis der Zwischenprüfung dar.

2.3. Ergebnisse

Insgesamt konnten die Daten von 51 Personen in die Analyse integriert werden. Die Ergebnisse der Abschlussprüfung liegen bei 48 Bewerbern vor, die der Zwischenprüfung lediglich bei 34. Zunächst werden die deskriptiven Statistiken der Prädiktoren und Kriterien betrachtet (vgl. Tabelle 1). Der Großteil der eingestellten Bewerberinnen und Bewerber weist gute bis sehr gute Empfehlungsgrade auf. Analog dazu ist der Gesamtpunktwert im Mittel mit etwa 107 Punkten überdurchschnittlich. Die Ergebnisse der Abschlussprüfungen sind mit einem Mittelwert von 9 Punkten, was der Schulnote 3 entspricht, als durchschnittlich zu bezeichnen. Die Ergebnisse der Zwischenprüfung liegen mit knapp 10 Punkten leicht darüber.

Tabelle 1: Deskriptive Statistiken der Prädiktoren und Kriterien

Variable	Mittelwert	Streuung	Minimum	Maximum
Empfehlungsgrad	4.57	0.63	2.50	5.00
Gesamtpunktwert	106.76	4.86	97.00	121.00
Punktwert der Abschlussprüfung	9.00	2.34	3.50	13.00
Punktwert der Zwischenprüfung	9.90	1.49	7.00	13.28

Anmerkung: Minimum und Maximum sind nicht die kleinsten und größten Werte der Skalen, sondern beziehen sich auf die Daten der Stichprobe.

Um den Zusammenhang zwischen den Prädiktoren und den Kriterien zu untersuchen, wird zunächst eine Kreuztabelle betrachtet (vgl. Tabelle 2). Alle Bewerberinnen und Bewerber, die ihre Abschlussprüfung mit der Note 2 bestanden haben, wurden von der DGP als (weitgehend) geeignet empfohlen. Dies trifft auch auf den Großteil der Bewerberinnen und Bewerber zu, die mit der Note 3 bestanden haben; nur für zwei Teilnehmer wurde eine mäßige Eignung prognostiziert. Insgesamt zeigt sich, dass Bewerberinnen und Bewerber mit schlechteren Leistungen in der Abschlussprüfung häufig auch geringere Empfehlungsgrade aufweisen. Zudem wird aus Tabelle 2 ersichtlich, dass die Durchschnittsnoten bei steigenden Empfehlungsgraden tendenziell höher sind als bei niedrigen.

Tabelle 2: Kreuztabelle DGP-Empfehlungsgrad und Ergebnis der Abschlussprüfung (Schulnote)

		DGP-Empfehlungsgrad					
		2.5	3.0	3.5	4.0	4.5	5.0
Ergebnis der Abschlussprüfung (Schulnote)	2	-	-	-	10 %	-	90 %
	3	-	4 %	4 %	12 %	16 %	64 %
	4	9 %	-	9 %	36 %	18 %	27 %
	5	-	50 %	-	50 %	-	-
Durchschnittsnote		4.0	4.0	3.5	3.6	3.3	2.8

Anmerkung: Die beiden betrachteten Variablen sind unterschiedlich gepolt. Der Wert 5 ist die schlechteste Leistung in der Abschlussprüfung, jedoch der höchste Empfehlungsgrad.

Die Validitätskoeffizienten sowohl des Empfehlungsgrades als auch des Gesamtpunktwertes sind außerordentlich gut (vgl. Tabelle 3). Sie erlauben gute Vorhersagen der Ergebnisse der Zwischenprüfung sowie der Abschlussprüfung. Der Zusammenhang der Prädiktoren zur Zwischenprüfung fällt dabei ein wenig höher aus als zur Abschlussprüfung. Dies ist auf Grund der geringeren Zeitspanne zwischen Eignungstest und Zwischenprüfung durchaus plausibel, sollte jedoch vor dem Hintergrund der recht geringen Fallzahl nicht überinterpretiert werden. Alle Koeffizienten sind auf dem Niveau von $\alpha = .01$ statistisch signifikant. Die Wahrscheinlichkeit, die vorliegenden Korrelationen unter der Nullhypothese ($r^* = 0$) zu finden, ist also niedriger als 1%.

Tabelle 3: Zusammenhang zwischen Vorhersagemaßen (Empfehlungsgrad / Gesamtpunktwert A2) und Ausbildungserfolg (Abschlussnote / Note der Zwischenprüfung)

Prädiktor	Kriterium	N	Validität	Validität (korrigiert)[1]
Empfehlungsgrad	Abschlussnote	48	r = .42 **	r = .69
Empfehlungsgrad	Zwischenprüfung	34	r = .44 **	r = .71
Gesamtpunktwert A2	Abschlussnote	48	r = .44 **	r = .58
Gesamtpunktwert A2	Zwischenprüfung	34	r = .52 **	r = .66

[1] Korrektur gegen Einschränkung der Prädiktorvarianz.
** *Die Korrelation ist auf dem Niveau von α = .01 statistisch signifikant.*

Wie bei Validitätsuntersuchungen üblich, ist die Varianz der Prädiktoren stark eingeschränkt: Ungeeignete Bewerberinnen und Bewerber erhalten seltener einen Ausbildungsplatz und sind der Analyse somit nicht zugänglich. Da niedrige Varianzen zu einer Unterschätzung der wahren Zusammenhänge führen (Stelzl, 2005), werden die Validitätskoeffizienten in der Regel gegen Einschränkungen der Prädiktorvarianz korrigiert. Ein Vergleich der korrigierten und unkorrigierten Werte verdeutlicht, dass insbesondere die Varianz des Empfehlungsgrades im Vergleich zur „wahren Varianz“ stark eingeschränkt ist.
Durch die Korrektur lassen sich die ermittelten Werte grob mit denen aus der wissenschaftlichen Literatur vergleichen. In einer Meta-Analyse, in der die Forschung aus 85 Jahren zusammengefasst wird, ermittelten Schmidt und Hunter (1998) einen korrigierten Zusammenhang zwischen Kognitiven Leistungstests und Ausbildungserfolg von r = .56. Die hier ermittelten Koeffizienten liegen zwischen r = .58 und r = .69 und entsprechen damit in etwa den in der Literatur berichteten Werten.

3. Ausblick

Die Ergebnisse verdeutlichen, dass der A2 in der Lage ist, den Ausbildungserfolg über mehrere Jahre valide vorherzusagen. Kann man daraus nun schließen, dass wir den A2 in dieser Form über Jahre weiterverwenden sollten, gemäß dem Motto „never change a running system“? Der Schluss liegt nahe, ist aber falsch.
Auch gute Testverfahren müssen regelmäßig aktualisiert und nachgebessert werden. Dies betrifft sowohl die Zusammensetzung des gesamten Tests aus den einzelnen Subtests als auch die Zusammensetzung der Subtests aus den Prüfungsfragen: manche Subtests messen die dahinterliegenden Konstrukte exakter als andere oder sagen die Leistung in der Abschlussprüfung besser vorher. Vielleicht ist eine Prüfungsfrage zu leicht oder zu schwer. Vielleicht ist die Bearbeitungszeit einzelner Subtests zu lang oder zu kurz.
Seit dem ersten Einsatz des A2 im Jahre 2006 konnten wir eine große Menge Daten sammeln, die uns hilft, genau diese Fragen zu beantworten und den A2 dahingehend noch weiter zu verbessern. Die neue Version kommt ab der Saison 2012 / 2013 unter dem Namen „A3“ zum Einsatz.

Literatur

Schmidt, F., & Hunter, J. (1998). The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings. Psychological Bulletin, 124(2), 262-274.

Stelzl, I. (2005). Fehler und Fallen der Statistik: für Psychologen, Pädagogen und Sozialwissenschaftler. Münster: Waxmann.

Korrespondenzanschriften der Autoren:

Dipl.-Psych. Benjamin Haarhaus
Deutsche Gesellschaft für Personalwesen e.V.
Grafenberger Allee 32
40237 Düsseldorf
haarhaus@dgp.de

Dipl.-Psych. Dr. Stephan Buchhester
Deutsche Gesellschaft für Personalwesen e.V.
Grassistr. 12
04107 Leipzig
buchhester@dgp.de

Dipl.-Psych. Nina Ristel
Deutsche Gesellschaft für Personalwesen e.V.
Stammestr. 40D
30459 Hannover
ristel@dgp.de