



Kompetenzzentrum für Bildungsevaluation und Leistungsmessung an der Universität Zürich · KBL
Centre de compétences en évaluation des formations et des acquis à l'Université de Zurich · CEA
Competence Centre for Educational Evaluation and Assessment at the University of Zurich · CEA

Wie werden die Ergebnisse in den Stellwerk-Tests interpretiert?

Von den Testergebnissen zu einer professionellen Beurteilung
der Kompetenzen der Schülerinnen und Schüler

Urs Moser

Zürich, September 2006

INHALT

1. Einleitung	3
2. Adaptives Testsystem.....	3
3. Testergebnisse interpretieren: Sozialer Vergleich	5
4. Testergebnisse interpretieren: Förderorientierter Vergleich	9
5. Testergebnisse interpretieren: Ein konkretes Beispiel	14
6. Von den Testergebnissen zur Stellwerk-Skala	16

1. Einleitung

Stellwerk ist die Bezeichnung für ein adaptives, webbasiertes, multimediales Testsystem. Stellwerk bietet Leistungstests für die Schülerinnen und Schüler der obligatorischen Schulzeit. Seit dem Frühjahr 2006 ist Stellwerk 8 auf dem Internet verfügbar. Stellwerk 8 umfasst verschiedene Leistungstests, die sich an den Lehrplänen und Lehrmitteln der 8. Klassen der Deutschschweiz orientieren. Im Frühjahr 2008 soll Stellwerk 9 verfügbar sein.

Die vorliegende Broschüre enthält die wichtigsten Informationen zum Testsystem. Insbesondere zeigt sie auf, wie die Testergebnisse zustande kommen und wie die Testergebnisse professionell genutzt und interpretiert werden. Die Informationen sind grösstenteils genereller Art und treffen für sämtliche Leistungstests von Stellwerk zu. Einige Angaben beziehen sich jedoch spezifisch auf Stellwerk 8.

Im Text wird unterschieden zwischen (1) Informationen, die sich auf die Oberfläche des Testsystems beziehen, und (2) Informationen, die einen vertieften Einblick ins Testsystem gewähren. Erstere sind für den richtigen Umgang mit den Ergebnissen unerlässlich und im Text blau unterlegt.

2. Adaptives Testsystem

Wie funktioniert ein adaptiver Test?

Stellwerk ist ein adaptives Testsystem. Das bedeutet, dass sich ein Test fortwährend den Fähigkeiten der Person anpasst, die den Test löst. Wie ist dies zu verstehen?

Ein Test beginnt für alle Personen mit einer relativ einfachen, zufällig ausgewählten Aufgabe. Nachdem die Aufgabe gelöst wurde, schätzt das System aus dem Schwierigkeitsparameter der Aufgabe und der Lösung (richtig oder falsch) die Fähigkeit der Person (Personenparameter). Danach sucht das System jene Aufgabe, deren Schwierigkeitsparameter am nächsten bei der geschätzten Fähigkeit der Person liegt. Löst beispielsweise eine Person alle Aufgaben von Beginn an richtig, dann schlägt sich dies in der Schätzung ihrer Fähigkeit nieder. Der Personenparameter wird grösser und dementsprechend weist das System der Person schwierigere Aufgaben zu. Umgekehrt sinkt der Personenparameter, wenn die Person die Aufgaben falsch löst. Das System weist der Person in diesem Fall Aufgaben zu, deren Schwierigkeitsparameter kleiner sind. Der Test dauert so lange, bis grössere Schwankungen bei der Schätzung der Fähigkeit ausbleiben und das System nur noch Aufgaben zuweist, deren Schwierigkeitsparameter sich kaum mehr unterscheiden. Die letzte Schätzung des Personenparameters entspricht dem Gesamtwert im Test.

Adaptive Tests haben den Vorteil, dass sie sich relativ rasch den Fähigkeiten der Schülerinnen und Schüler anpassen und zu einem sehr zuverlässigen Testergebnis führen. Schwache Schülerinnen und Schüler werden nicht mit zu schwierigen Aufgaben frustriert, starke werden nicht mit zu einfachen Aufgaben gelangweilt.

Aus wie vielen Aufgaben bestehen die Stellwerk-Tests?

Damit ein adaptives Testsystem wunschgemäss funktioniert, muss es auf eine genügend grosse Anzahl von Aufgaben zurückgreifen können. Dazu wird eine sogenannte Itembank entwickelt. In der Itembank werden erprobte und bewährte Testaufgaben gesammelt, deren Schwierigkeit bekannt ist. Die gegenwärtige Anzahl Aufgaben der Tests von Stellwerk 8 sind in Tabelle 1 für jeden geprüften Fachbereich und Teilbereich dargestellt. Die einzelnen Tests werden laufend mit neuen Aufgaben ergänzt.

Tabelle 1: Anzahl Aufgaben in der Itembank nach Fachbereich und Teilbereich

Fachbereiche	Teilbereiche	331
Mathematik		
	Zahlen, Grössen, Operationen (Arithmetik)	156
	Form und Mass in Ebene und Raum (Geometrie)	86
	Variable, Term, Gleichung (Algebra)	48
	Datendarstellung, Datenanalyse und Zufall, funktionale Zusammenhänge und ihre Darstellungsformen (Stochastik, Funktionen)	41
Deutsch		
	Hören	83
	Lesen	78
	Sprachreflexion und Rechtschreibung	91
Französisch		
	Hören	77
	Lesen	75
Englisch		
	Hören	86
	Lesen	107
Natur und Technik		
	Biologie	51
	Chemie	49
	Physik	50
TOTAL		1078

3. Testergebnisse interpretieren: Sozialer Vergleich

Beurteilung und Interpretation von Testergebnissen mit Hilfe von sozialen Vergleichsnormen

Die einfachste Form, Testergebnisse von Schülerinnen und Schülern auszuweisen, ist, die Anzahl richtig gelöster Aufgaben zu einer Punktzahl zusammenzuzählen und diese zusätzlich in den Prozentanteil richtig gelöster Aufgaben zu transformieren. Der Vorteil dieser einfachen Darstellung von Testergebnissen ist, dass sie allgemein verständlich ist. Auch für die Schülerinnen und Schüler ist nachvollziehbar, wie das Testergebnis zustande gekommen ist.

Die Anzahl Punkte oder der Prozentanteil richtig gelöster Aufgaben sagen allerdings noch nichts darüber aus, ob ein Testergebnis als gut oder als schlecht zu beurteilen ist. Es lässt sich nicht beurteilen, ob beispielsweise 30 Prozent richtig gelöster Aufgaben einem eher guten oder einem eher schlechten Ergebnis entsprechen. Testergebnisse können erst dann beurteilt und interpretiert werden, wenn Angaben darüber vorliegen, wie gut der Test in einer Population gelöst wird. Aus diesem Grund werden Tests in der Regel in einer Referenzpopulation normiert, was nichts anderes bedeutet, als dass die Tests von Mitgliedern einer Population gelöst werden, um anschliessend die Verteilung der Testergebnisse in dieser Population zu berechnen.

Eine Population ist eine eindeutig definierte Gruppe von Individuen der Bevölkerung, über die etwas ausgesagt werden soll, beispielsweise die stimmberechtigte Bevölkerung der Deutschschweiz, alle 15-Jährigen Europas oder alle Schülerinnen und Schüler der 8. Klassen des Kantons St. Gallen¹.

Weil es in vielen Fällen zu aufwändig wäre, eine ganze Population zu testen, werden in der Regel repräsentative Stichproben aus der Population getestet. Sofern eine genügend grosse Stichprobe² nach wissenschaftlichen Kriterien gebildet wird und die Stichprobe damit einem repräsentativen Abbild der Population entspricht, wird von der Verteilung der Testergebnisse in der Stichprobe auf die Verteilung der Testergebnisse in der Population geschlossen (schliessende Statistik). Dieser Prozess wird üblicherweise als Normierung oder Eichung bezeichnet.

¹ Die Definition bezieht sich auf den vorliegenden spezifischen Kontext. Eine Population oder Grundgesamtheit bezieht sich allgemein auf die Gesamtheit aller gleichartigen Elemente oder Objekte, über die etwas ausgesagt werden soll.

² Bei einer Meinungsumfrage werden in der Regel mindestens 500 zufällig ausgewählte Personen befragt, damit von den Ergebnissen der Stichprobe zuverlässig auf die Population geschlossen werden kann. Im Bildungsbereich braucht es für eine gleich zuverlässige Schätzung von Parametern aus der Stichprobe auf Parameter in der Population mehr als 500 Schülerinnen und Schüler, weil diese in Klassen unterrichtet werden und deshalb nicht voneinander unabhängig sind. Die Fähigkeiten von Schülerinnen und Schülern, die die gleiche Klasse besuchen, sind sich in der Regel ähnlicher als die Fähigkeiten von Schülerinnen und Schülern, die unterschiedliche Klassen besuchen.

Mit der Normierung oder Eichung wird ein Bezugssystem für die Einordnung der individuellen Testergebnisse geschaffen. Zum Beispiel wird aufgezeigt, welcher Prozentanteil der Population wie viele Testaufgaben richtig löst. Diese Angaben ermöglichen es, individuelle Testergebnisse mit der Verteilung der Testergebnisse in der Population zu vergleichen und zu beurteilen. Ein solcher Vergleich könnte zu folgender Erkenntnis führen: Wenn der durchschnittliche Anteil richtig gelöster Aufgaben in einer Population bei 70 Prozent liegt, dann entsprechen 30 Prozent richtig gelöster Aufgaben einem vergleichsweise schlechten Ergebnis. Liegt der durchschnittliche Anteil richtig gelöster Aufgaben in der Population hingegen bei 15 Prozent, dann sind 30 Prozent richtig gelöster Aufgaben ein vergleichsweise gutes Ergebnis.

Mit welchen Schülerinnen und Schülern wurden die Stellwerk-Tests normiert?

Die Normierung der Tests von Stellwerk 8 wurde im Frühling 2005 mit allen Schülerinnen und Schülern der 8. Klassen des Kantons St. Gallen durchgeführt. Dass Tests mit einer Population normiert werden, ist eher selten. Populationen sind meist sehr gross, weshalb auf eine Erhebung bei allen Mitgliedern der Population verzichtet und die Normierung mit Hilfe einer repräsentativen Stichprobe aus der Population durchgeführt wird. Weil für Stellwerk 8 sehr viele Testaufgaben normiert werden mussten, war der Einbezug aller Schülerinnen und Schüler der 8. Klassen des Kantons St. Gallen (Referenzpopulation) jedoch sinnvoll. Insgesamt konnten dank der Teilnahme von über 6000 Schülerinnen und Schülern mehr als 1000 Testaufgaben normiert beziehungsweise geeicht werden.

Wie werden die Testergebnisse in den Stellwerk-Tests ausgewiesen?

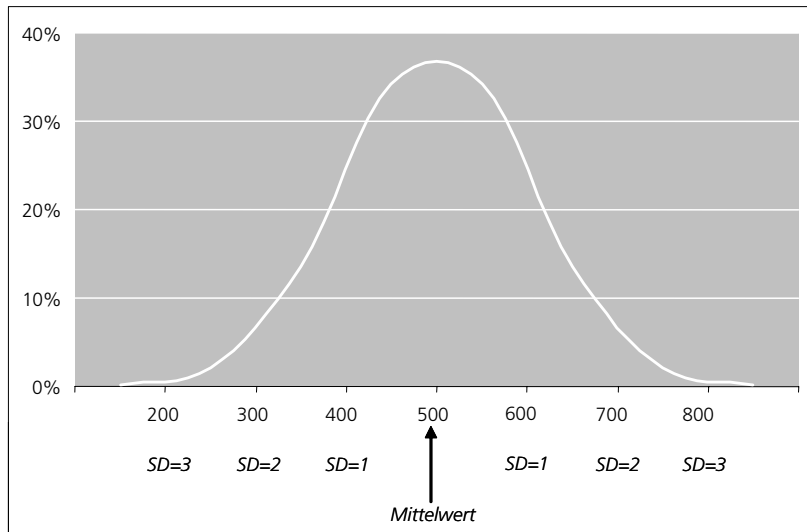
Beim adaptiven Testen bearbeiten die Schülerinnen und Schüler individuelle Tests, die unterschiedlich schwierig sind. Dies hat zur Folge, dass die Anzahl oder der Prozentanteil richtig gelöster Aufgaben zweier Schülerinnen und Schüler nicht sinnvoll miteinander verglichen werden kann. Auch der Vergleich des Anteils richtig gelöster Aufgaben mit den Testergebnissen der normierten Referenzpopulation hilft nicht weiter.

Weil die Anzahl oder der Anteil richtig gelöster Aufgaben bei einem adaptiven Testsystem zu keiner interpretierbaren Grösse führt, werden die individuellen Ergebnisse auf einer standardisierten Skala – der Stellwerk-Skala – ausgewiesen. Die Stellwerk-Skala hat – wie beispielsweise die PISA-Skala – einen Mittelwert von 500 Punkten und eine Standardabweichung von 100 Punkten³.

Leistungstests sind meist so konstruiert, dass sich die Verteilung der Testergebnisse einer Normalverteilungskurve annähert (vgl. Abbildung 1). Die Normalverteilungskurve hat die Form einer Glockenkurve. Das bedeutet, dass die mittleren Werte häufig, die extrem hohen oder tiefen Werte selten vorkommen.

³ **Achtung!** Die Stellwerk-Skala und die PISA-Skala haben zwar die gleiche Metrik, sie sind aber ansonsten nicht vergleichbar. Zum einen basieren die Tests auf verschiedenen Referenzsystemen. Zum andern sind die PISA-Tests für 15-Jährige der OECD-Länder, die Stellwerk-Tests für Schülerinnen und Schüler der 8. Klasse des Kantons St. Gallen normiert.

Abbildung 1: Normalverteilung



Soll die Beurteilung und Interpretation eines Stellwerk-Testergebnisses im sozialen Vergleich erfolgen, dann sind folgende Eigenschaften der Normalverteilung hilfreich. Rund 68 Prozent der Testergebnisse liegen zwischen 400 und 600 Punkten (Mittelwert $[M = 500] \pm$ eine Standardabweichung $[1 \text{ SD} = 100]$), rund 95 Prozent der Testergebnisse liegen zwischen 300 und 700 Punkten (Mittelwert $[M = 500] \pm$ zwei Standardabweichungen $[2 \text{ SD} = 200]$) und nahezu alle Testergebnisse liegen zwischen 200 und 800 Punkten (Mittelwert $[M = 500] \pm$ drei Standardabweichungen $[3 \text{ SD} = 300]$).

Wie entsprechen sich die Ergebnisse in den Stellwerk-Tests und Prozentränge?

Jeder Punktwert auf der Stellwerk-Skala kann auch in einen Prozentrang transformiert werden. Der Prozentrang entspricht dem Prozentanteil der Personen, die ein bestimmtes Testergebnis erreichen. Ein Prozentrang von 10 bedeutet beispielsweise für eine Schülerin, dass 9 Prozent der Population ein tieferes Testergebnis, 90 Prozent ein höheres Testergebnis erreichen. Tabelle 2 zeigt, welche Punktzahl auf der Stellwerk-Skala welchen Prozenträngen entspricht.

Die Beziehung zwischen der Prozentrang-Skala und der Stellwerk-Skala zeigt, dass die mittleren Werte auf der Stellwerk-Skala (zwischen 400 und 600 Punkten) häufig vorkommen, die hohen Werte (zwischen 700 und 800 Punkten) oder die tiefen Werte (zwischen 300 und 200 Punkten) eher selten vorkommen. Ein Prozentrangwerte-Unterschied zwischen 90 und 95 Prozent bedeutet demnach in Punkten auf der Stellwerkskala viel mehr als ein Prozentrangunterschied zwischen 50 und 55 Punkten⁴.

⁴ Eine gute Einführung in die Interpretation von standardisierten Testwerten geben beispielsweise G.A. Lienert & A. von Eye. *Erziehungswissenschaftliche Statistik. Eine elementare Einführung für pädagogische Berufe*. Basel: Beltz. 1994.

Tabelle 2: Beziehung zwischen der Prozentrang-Skala und der Stellwerk-Skala

Prozentrang-Skala	Stellwerk-Skala
0.1%	200
1%	229
2.3%	300
5%	336
10%	372
15%	397
15.9%	400
20%	416
25%	433
30%	448
35%	462
40%	475
45%	488
50%	500
55%	513
60%	525
65%	539
70%	553
75%	568
80%	584
84.1%	600
85%	604
90%	628
95%	665
97.7%	700
99%	771
99.9%	800

Was heisst es, wenn mehrere Schülerinnen und Schüler einer Klasse einen Gesamtwert von 200 oder 800 Punkten erreichen?

Es ist möglich, allerdings eher selten, dass gleich mehrere Schülerinnen und Schüler einer Klasse einen Gesamtwert erreichen, der bei 200 oder 800 Punkten liegt. Zwei Schülerinnen und Schüler einer Klasse von 20 Schülerinnen und Schülern entsprechen bereits 10 Prozent der Klasse. Bezogen auf die Referenzpopulation (alle Schülerinnen und Schüler der 8. Klassen des Kantons St. Gallen) entsprechen zwei Schülerinnen und Schüler allerdings nur einem Anteil von 0.03 Prozent. Die Verteilung der Testergebnisse innerhalb einer Klasse kann von der Verteilung der Testergebnisse innerhalb der Population stark abweichen. Diese Abweichung ist nicht falsch, sondern sie sagt einzig etwas über die Kompetenzen der Schülerinnen und Schüler aus, die im Vergleich zur Referenzpopulation sehr tief oder sehr hoch sind.

4. Testergebnisse interpretieren: Förderorientierter Vergleich

Beurteilung und Interpretation von Testergebnissen mit Hilfe von konkreten Angaben zu den Kompetenzen

Den beiden bisher vorgestellten Kennzahlen – Testergebnisse als Anteil richtig gelöster Aufgaben und Testergebnisse in standardisierter Form – ist eines gemeinsam: Die Interpretation der Testergebnisse erfolgt primär über den sozialen Vergleich. Ob ein Testergebnis als gut oder schlecht beurteilt wird, ergibt sich aus der Stellung des Testergebnisses innerhalb der Population. Aufgrund der Normierung eines Testergebnisses kann genau festgestellt werden, wo ein individuelles Testergebnis im Vergleich zu den Testergebnissen aller Schülerinnen und Schüler liegt. Diese Art der Interpretation erlaubt allerdings nur beschränkt Aussagen darüber, was eine Schülerin oder ein Schüler mit einem bestimmten Testergebnis konkret weiss oder kann. Was fehlt, ist eine Umschreibung der Kompetenzen, die einem bestimmten Testergebnis (Anteil richtig gelöster Aufgaben oder standardisierter Wert) entspricht.

Damit die Testergebnisse von Leistungsmessungen zur Förderung der Schülerinnen und Schüler genutzt werden können, müssen sie auch in Bezug zu klar umschriebenen Kompetenzen interpretiert werden können. Der Bezug zwischen Testergebnis und Kompetenz wird bei Stellwerk sowohl mit Deskriptoren und Begriffen, auf den sich ein Test bezieht, als auch mit Testaufgaben hergestellt. Deskriptoren, Begriffe und Beispielaufgaben sind im Referenzrahmen und in den Interpretationshilfen ausgewiesen.

Mit der Normierung oder Eichung des Tests wird nicht nur ein Bezugssystem für die Einordnung der individuellen Testergebnisse im sozialen Vergleich geschaffen, sondern zugleich auch ein Bezugssystem für die Einordnung der individuellen Testergebnisse im Vergleich zur ermittelten Kompetenz⁵.

Aus diesem Grund wurde für jeden Fachbereich (Mathematik, Deutsch, Englisch, Französisch sowie Natur und Technik) und für jeden Teilbereich (beispielsweise Arithmetik oder Hörverstehen) ermittelt, wie die Testergebnisse der Schülerinnen und Schüler mit den im Referenzrahmen beschriebenen Kompetenzen und den im Test eingesetzten Aufgaben zusammenhängen. Doch wie ist das überhaupt möglich?

Wie die meisten modernen Tests basieren auch die Stellwerk-Tests auf der Annahme, dass die Wahrscheinlichkeit für einen Schüler, eine Testaufgabe richtig zu lösen, grundsätzlich von zwei Merkmalen abhängt: (1) von seiner Fähigkeit und (2) von der Aufgabenschwierigkeit. Je grösser die Fähigkeit ist, desto wahrscheinlicher ist es, dass ein Schüler eine be-

⁵ Bei der Beurteilung der Schülerinnen und Schüler wird in der Regel zwischen sozialer Bezugsnorm, kriterialer Bezugsnorm und individueller Bezugsnorm unterschieden. Die Leistungen eines Individuums werden in der Schule entweder im Vergleich zu den Mitschülerinnen und Mitschülern beurteilt (sozialer Vergleich), im Vergleich zu einem Kriterium wie beispielsweise ein Lehrplanziel (kriterialer Vergleich) oder im Vergleich zum individuellen Fortschritt (individueller Vergleich).

stimmte Aufgabe richtig löst. Und je schwieriger eine Aufgabe ist, desto unwahrscheinlicher ist es, dass ein Schüler mit einer bestimmten Fähigkeit die Aufgabe richtig löst⁶.

Sofern von dieser Annahme beziehungsweise von diesem Modell ausgegangen wird, muss für jede Testaufgabe bestimmt werden, wie gross die Lösungswahrscheinlichkeit bei einer bestimmten Fähigkeit des Schülers ist. Das heisst, die Schwierigkeiten der Aufgaben (Schwierigkeitsparameter) und die Testergebnisse der Schülerinnen und Schüler (Personenparameter) werden berechnet. Diese Parameterschätzung ist beim angewendeten Modell kompliziert, weil es keine expliziten Gleichungen beziehungsweise Formeln gibt, die jeweils nach einer unbekanntem Grösse auflösbar sind⁷. Ziel dieser Berechnungen ist es, die Wahrscheinlichkeit für die Schülerinnen und Schüler, dass sie eine bestimmte Aufgabe richtig lösen, als Funktion ihrer Fähigkeiten zu berechnen.

Der grosse Vorteil dieser Methode besteht darin, dass sowohl das Testergebnis der Schülerinnen und Schüler (Personenparameter) als auch die Schwierigkeiten der Aufgaben (Schwierigkeitsparameter) auf der gleichen Skala – der Stellwerk-Skala – abgebildet werden. Zwischen der Fähigkeit eines Schülers und der Schwierigkeit einer Aufgabe besteht eine Beziehung, die für die Interpretation der Testergebnisse genutzt werden kann. Dadurch kann ein individuelles Testergebnis nicht nur als absolute Punktzahl mit Hilfe des sozialen Vergleichs interpretiert werden, sondern in Bezug zu den Testaufgaben und zum Referenzrahmen, der die Basis für die Entwicklung der Testaufgaben bildet. Konkret heisst dies, dass sich aufgrund des individuellen Testergebnisses für jede Aufgabe bestimmen lässt, wie wahrscheinlich es ist, dass ein Schüler oder eine Schülerin die Aufgabe richtig löst.

⁶ Nun mag es aussergewöhnlich erscheinen, dass die erreichte Punktzahl in einem Test mit Lösungswahrscheinlichkeiten für die einzelnen Testaufgaben in Verbindung gebracht wird. Wahrscheinlichkeitsangaben sind aber immer dann angebracht, wenn eine Aussage nicht mit hundertprozentiger Sicherheit gemacht werden kann. Und dies gilt auch für ein Testergebnis. Ein Testergebnis führt zwar zu einer objektiven und zuverlässigen Information. Aufgrund eines Testergebnisses kann aber nicht mit letzter Sicherheit ausgesagt werden, was eine Schülerin oder ein Schüler kann. Ein anderes Beispiel für Wahrscheinlichkeitsaussagen finden wir bei den Wetterprognosen. Wetterprognosen sind ebenfalls mit einem Unsicherheitsfaktor behaftet, weshalb sie in der Regel als Wahrscheinlichkeitsaussagen angegeben werden: Die Prognose für den nächsten Tag trifft meistens – oder als Wahrscheinlichkeitssaussage mit 90 Prozent Sicherheit –, aber nicht in jedem Fall zu. Die Prognose für das Wetter des übernächsten Tags trifft mit einer geringeren Sicherheit, beispielsweise mit 85 Prozent Sicherheit, zu.

⁷ Die Berechnung der Parameter beziehungsweise der Testergebnisse wird an dieser Stelle nicht weiter ausgeführt. Sie ist im Lehrbuch von Jürgen Rost. *Lehrbuch Testtheorie Testkonstruktion*. Bern: Verlag Hans Huber, 2004, ausführlich beschrieben.

Wie wird die Beziehung zwischen Fähigkeit und Aufgabenschwierigkeit interpretiert?

Bei Stellwerk lässt sich sowohl das Testergebnis einer Schülerin (Personenparameter) als auch die Schwierigkeit einer Aufgabe (Schwierigkeitsparameter) auf der Stellwerk-Skala von 200 bis 800 Punkten darstellen. Die Stellwerk-Skala wurde so gebildet, dass bei Übereinstimmung der Fähigkeit eines Schülers mit der Schwierigkeit einer Aufgabe – das heißt, Personenparameter und Schwierigkeitsparameter sind identisch – der Schüler die Aufgabe mit einer Wahrscheinlichkeit von 62 Prozent ($p = 0.62$) richtig löst.

Erreicht ein Schüler beispielsweise im Mathematiktest 300 Punkte auf der Stellwerk-Skala, dann hat er bei einer Aufgabe, deren Schwierigkeit ebenfalls mit 300 Punkten auf der Stellwerk-Skala angegeben ist, eine Lösungswahrscheinlichkeit von 62 Prozent. Löst der Schüler eine einfachere Aufgabe, beispielsweise eine Aufgabe mit einem Schwierigkeitsparameter von 250 Punkten, dann steigt die Lösungswahrscheinlichkeit auf 82 Prozent. Löst der Schüler eine schwierigere Aufgabe, beispielsweise eine Aufgabe mit einem Schwierigkeitsparameter von 350 Punkten, dann sinkt die Lösungswahrscheinlichkeit auf 50 Prozent.

Weshalb wird die Beziehung zwischen den Fähigkeiten der Schülerinnen und Schüler und den Schwierigkeiten der Aufgaben exakt über die Lösungswahrscheinlichkeit von $p = 0.62$ dargestellt?

Eigentlich lässt sich dies frei bestimmen. Bei Stellwerk wurde der Wert von $p = 0.62$ aufgrund der Beschreibung der Kompetenzen in der Interpretationshilfe festgelegt. Die Interpretationshilfe enthält für jeden Kompetenzbereich eine Zuteilung von ausgewählten Aufgabenbeispielen und Kompetenzbeschreibungen auf der Stellwerk-Skala. Die Stellwerk-Skala ist in sechs Intervalle unterteilt (200–300 Punkte; 301–400 Punkte; 401–500 Punkte; 501–600 Punkte; 601–700 Punkte; 701–800 Punkte). Jedes Intervall ist durch typische Aufgabenbeispiele (blauer Text in der Interpretationshilfe) illustriert, deren Schwierigkeitsparameter jeweils im entsprechenden Intervall liegen. Eine Aufgabe mit dem Schwierigkeitsparameter von 750 Punkten auf der Stellwerk-Skala wird in der Interpretationshilfe dem Intervall von 700 bis 800 Punkten zugeordnet. Ein Intervall enthält aber auch Kompetenzbeschreibungen beziehungsweise Deskriptoren (schwarzer Text in der Interpretationshilfe) oder Begriffe (roter Text in der Interpretationshilfe), die für die Lösung von Aufgaben mit entsprechenden Schwierigkeitsparametern vorausgesetzt werden.

Die gewählte Lösungswahrscheinlichkeit von 62 Prozent ($p = 0.62$) hat zur Folge, dass beispielsweise eine Schülerin, deren Testergebnis am unteren Ende eines Intervalls liegt, eine durchschnittliche Lösungswahrscheinlichkeit der Aufgaben des Intervalls von $p = 0.5$ hat. Erreicht die Schülerin beispielsweise im Mathematiktest 300 Punkte, dann beträgt die durchschnittliche Lösungswahrscheinlichkeit für alle Aufgaben des Intervalls (zwischen 300 und 400 Punkten) $p = 0.5$; das heißt, sie löst vermutlich 50 Prozent der Aufgaben dieses Intervalls richtig. Tabelle 3 zeigt, wie sich die Lösungswahrscheinlichkeit für Schülerinnen und Schüler in Abhängigkeit ihrer Fähigkeiten (Personenparameter) und der Aufgabenschwierigkeiten (Schwierigkeitsparameter) verändert.

Tabelle 3: Lösungswahrscheinlichkeiten in Abhängigkeit von Testergebnissen und Aufgabenschwierigkeiten

Testergebnisse (Personenparameter)	Aufgabenschwierigkeiten (Schwierigkeitsparameter)		
	300	350	400
400	82%	73%	62%
350	73%	62%	50%
300	62%	50%	38%

Das Testergebnis in Form der Punktzahl auf der Stellwerk-Skala zeigt den Schülerinnen und Schülern mit Hilfe der Interpretationshilfe, über welche Kompetenzen und Begriffe sie mit einer mittleren Wahrscheinlichkeit verfügen beziehungsweise welche Aufgaben sie mit einer mittleren Wahrscheinlichkeit lösen können. Die Lösungswahrscheinlichkeit liegt je nach Fähigkeit innerhalb eines Intervalls zwischen 38 und 82 Prozent. Beispielsweise löst ein Schüler mit der Fähigkeit von 400 Punkten die meisten Aufgaben im Intervall zwischen 300 und 400 Punkten ohne Probleme. Kommt der Schüler nur auf 300 Punkte, hat er beim Lösen der schwierigeren Aufgaben meist noch Probleme. Aufgaben eines tieferen Intervalls, beispielsweise jene zwischen 200 und 300 Punkten, kann der Schüler hingegen mit sehr hoher Wahrscheinlichkeit richtig lösen. Aufgaben eines höheren Niveaus, beispielsweise zwischen 400 und 500 Punkten, kann der Schüler erst mit geringer Wahrscheinlichkeit lösen.

Wie werden die Ergebnisse für die Förderung der Schülerinnen und Schüler genutzt?

Testergebnisse als Wahrscheinlichkeitsaussagen sind ungewohnt. Die Angabe von Wahrscheinlichkeitsaussagen kennt man aus dem Alltag am ehesten aus der Wettervorhersage. Je langfristiger die Wettervorhersage ist, desto geringer ist die Wahrscheinlichkeit, dass sie zutrifft. Je kurzfristiger die Prognose ist, desto grösser ist die Wahrscheinlichkeit, dass sie zutrifft.

Bei Stellwerk gilt dieses Prinzip in ähnlicher Weise. Erreicht ein Schüler ein sehr hohes individuelles Testergebnis, so löst er mit an hundertprozentiger Sicherheit grenzender Wahrscheinlichkeit die sehr einfachen Testaufgaben, die zuunterst auf der Kompetenzskala angesiedelt sind. Selbstverständlich besteht immer eine gewisse Wahrscheinlichkeit, dass ein sehr guter Schüler aus irgendwelchen Gründen eine sehr einfache Aufgabe falsch löst.⁸

⁸ Mit steigender Schwierigkeit der Aufgaben nimmt die Lösungswahrscheinlichkeit etwas ab. Weil bei der Anwendung dieses Modells die Testergebnisse in Form von Wahrscheinlichkeitsaussagen gemacht werden, wird es manchmal auch probabilistisches Testmodell bezeichnet. Eine andere Bezeichnung für Tests, die auf dieser Annahme beruhen, ist die Bezeichnung Item Response Theory. Der Name rührt daher, dass zwischen einer Aufgabe (Item) und dem Antwortverhalten der Person (Response) eine Beziehung in Form einer Funktion dargestellt wird.

Nun wäre es für die Förderung im Unterricht wenig hilfreich, für die Interpretation der Testergebnisse nur einzelne Aufgaben beizuziehen. Förderung im Unterricht heisst nicht, Testaufgaben zu üben, sondern den Schülerinnen und Schülern mit geeigneten Unterrichtsarrangements die Möglichkeiten zu bieten, sich jene Begriffe und Kompetenzen anzueignen, die für das Lösen der Aufgaben notwendig sind. Aus diesem Grund sind die Intervalle in der Interpretationshilfe mit Kompetenzen in Form von Deskriptoren und Begriffen umschrieben. Das Lösen der Aufgaben eines Intervalls setzt jeweils das Beherrschen gleicher Kompetenzen und das Verständnis ähnlicher Begriffe voraus. Die Aufgaben eines Intervalls weisen deshalb auch vergleichbare Anforderungen und Merkmale auf, die sich zugleich von den Anforderungen und Merkmalen der Aufgaben der tiefer oder höher liegenden Intervalle unterscheiden.

Wie zuverlässig sind die Testergebnisse?

Stellwerk führt zu einer sehr präzisen Beurteilung der Kompetenzen in einem Fachbereich, weil ein Test erst dann abgebrochen wird, wenn das Lösen weiterer Aufgaben das Testergebnis kaum mehr verändert. Das Testergebnis wird im Zertifikat für jeden Fachbereich mit einem Gesamtwert ausgewiesen. Dieser Wert wird auf Grund von allen im Fachbereich bearbeiteten Testaufgaben berechnet. Er liegt zwischen 200 und 800 Punkten.

Selbstverständlich besteht auch die Möglichkeit, dass die Punktzahl kleiner als 200 oder grösser als 800 Punkte ist. Diese Ergebnisse sind aber eher selten (< als 1 Prozent), weshalb die Skala auf 200 bis 800 Punkte beschränkt wird. Zudem gibt es im Testsystem auch kaum Aufgaben, deren Schwierigkeitsparameter über 800 Punkten oder unter 200 Punkten liegen. Die Testergebnisse in den extrem tiefen oder extrem hohen Bereichen liessen sich gar nicht interpretieren, weil die Tests für diese Kompetenzen zu wenige oder keine Aufgaben aufweisen.

Die Ergebnisse in den Teilbereichen, beispielsweise «Zahlen und Zahlenraum» im Fachbereich Mathematik oder «Hören» im Fachbereich Englisch, führen zu einem Profil innerhalb eines Fachbereichs. Das Profil zeigt, in welchen Teilbereichen die Schülerinnen und Schüler besser oder weniger gut abgeschnitten haben. Die Ergebnisse in den Teilbereichen werden aufgrund jener Aufgaben berechnet, die dem entsprechenden Teilbereich angehören. Dabei handelt es sich immer nur um eine Teilmenge aller Aufgaben eines Tests. Die Ergebnisse in den Teilbereichen werden zudem erst nach Abbruch des Tests auf Grund des Anteils richtig gelöster Aufgaben und der durchschnittlichen Schwierigkeit der im Teilbereich gelösten Aufgaben berechnet; also nicht adaptiv wie der Gesamtwert des Fachbereichs. Dies hat zur Folge, dass die Ergebnisse in den Teilbereichen weniger präzise sind als das Gesamtergebnis im Fachbereich.

Trotzdem spiegeln die Ergebnisse in den Teilbereichen den Gesamtwert im Fachbereich in den meisten Fällen sehr gut. Es wäre allerdings falsch, den Gesamtwert aus den Ergebnissen in den Teilbereichen zu berechnen. Zum einen sind die Ergebnisse in den Teilbereichen etwas weniger präzise. Zum andern können sehr gute oder sehr schlechte Ergebnisse in einem Teilbereich zu sehr hohen oder sehr tiefen Werten im Teilbereich führen. Weil die

Stellwerk-Skalen am unteren und oberen Ende nicht linear verlaufen, führt das arithmetische Mittel der Ergebnisse in den Teilbereichen zu einer nicht interpretierbaren Grösse, die zum Teil erheblich vom Gesamtwert des Fachbereichs abweichen kann.

Wofür sollen welche Testergebnisse verwendet werden?

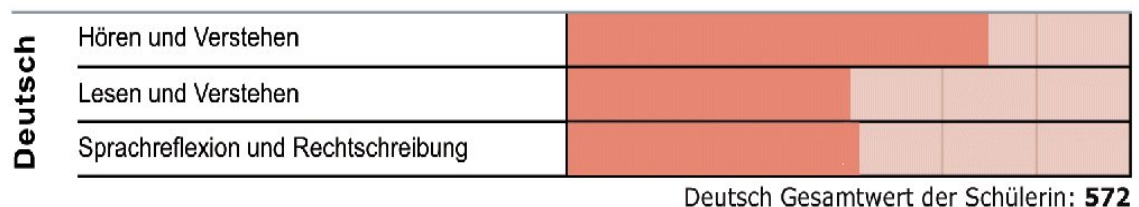
Für Beratungsgespräche mit Eltern oder Behörden, aber auch für Laufbahnentscheide werden mit Vorteil die Gesamtwerte der Fachbereiche als relevante Grössen genutzt. Für die Förderung im Unterricht und die Planung des Lehr-Lern-Prozesses der Schülerinnen und Schüler sollten zusätzlich die Ergebnisse in den Teilbereichen berücksichtigt werden.

5. Testergebnisse interpretieren: Ein konkretes Beispiel

Ergebnisse im Zertifikat

Abbildung 2 zeigt einen Ausschnitt aus einem Zertifikat, das eine Schülerin nach der Durchführung des Deutschtests erhalten hat. Die Schülerin hat im Fachbereich Deutsch einen Gesamtwert von 572 Punkten erreicht. Die Teilbereiche Sprechen und Schreiben werden nicht auf der Stellwerk-Skala ausgewiesen⁹.

Abbildung 2: Ausschnitt aus einem Zertifikat: Fachbereich Deutsch



Interpretationshilfen

Zur Interpretation der Testergebnisse im sozialen Vergleich können die Eigenschaften der Normalverteilung beigezogen werden (soziale Bezugsnorm). Zur Interpretation der Testergebnisse im förderdiagnostischen Vergleichen können die inhaltlichen Beschreibungen der Kompetenzen in den Interpretationshilfen genutzt werden. Diese befinden sich auf dem Internet (www.Stellwerk-Check.ch).

⁹ Sprechen und Schreiben lassen sich nur beschränkt adaptiv übers Internet testen, weshalb die Ergebnisse auch nicht auf der Stellwerk-Skala abgebildet werden können. Stellwerk plant aber, einzelne Schreibkompetenzen adaptiv übers Internet zu erfassen.

Wie gut ist das Testergebnis der Schülerin im sozialen Vergleich zu beurteilen?

Das Testergebnis (Gesamtwert von 572 Punkten) im Fachbereich Deutsch liegt deutlich über dem Mittelwert der Referenzpopulation (500 Punkte). Mit Hilfe der Angaben über den Zusammenhang zwischen der Stellwerk-Skala und dem Prozentrang (vgl. Tabelle 2) kann festgestellt werden, dass rund ein Viertel aller Schülerinnen und Schüler der Referenzpopulation (Schülerinnen und Schüler der 8. Klasse des Kantons St. Gallen) ein höheres Testergebnis erreichen. Rund drei Viertel der Referenzpopulation erreichen ein Testergebnis, das tiefer als 572 Punkte ist.

Der Vergleich der Ergebnisse in den Teilbereichen zeigt zudem, dass die Aufgaben im Teilbereich «Hören» besser gelöst werden als die Aufgaben in den Teilbereichen «Lesen» und «Sprachreflexion/Rechtschreibung». Die Schülerin verfügt im Teilbereich «Hören» über deutlich höhere Kompetenzen als in den übrigen beiden Teilbereichen.

Über welche Kompetenzen verfügt die Schülerin?

Im Teilbereich «Hören» liegt das Testergebnis leicht über 600 Punkten. Ein Blick auf die Interpretationshilfe für den Fachbereich Deutsch zeigt, dass dem Intervall zwischen 600 und 700 Punkten Kompetenzen wie beispielsweise das Unterscheiden von Textsorten (Rede, Interview, Erzählung, Reportage, Kommentar, Hörspiel) zugeordnet werden. Mit Hilfe der Angaben über die Lösungswahrscheinlichkeiten in Abhängigkeit von Testergebnissen und Aufgabenschwierigkeiten (vgl. Tabelle 3) kann festgestellt werden, dass die Schülerin mit mittlerer Wahrscheinlichkeit bereits Aufgaben lösen kann, die Anforderungen stellen, wie sie im Intervall zwischen 600 und 700 Punkten beschrieben sind. Einfachere Kompetenzen, beispielsweise die gehörte Information zur Beantwortung von Fragen nutzen (aus den Wetterprognosen Rückschlüsse auf die Durchführung von Anlässen ziehen), sind mit sehr hoher Wahrscheinlichkeit vorhanden.

Im Teilbereich «Lesen» liegt das Testergebnis bei rund 500 Punkten. Das heisst, dass die Kompetenzen im Intervall zwischen 400 und 500 Punkten bereits vorhanden sind. Fragen zum Text können beantwortet werden, Lernstrategien für die Zusammenfassung eines Textes sind bekannt (Textstellen markieren und Randnotizen erstellen). Etwas mehr Mühe bereiten die Aufgaben des Intervalls zwischen 500 und 600 Punkten, beispielsweise das Sammeln von verschiedenen Informationen zur Vorbereitung eines Vortrags.

Das Testergebnis im Teilbereich «Sprachreflexion und Rechtschreibung» liegt ebenfalls bei rund 500 Punkten. Das heisst, dass die verbalen Teile eines Satzes mit mittlerer Wahrscheinlichkeit bestimmt werden können. Die Gross- und Kleinschreibung bereitet der Schülerin hingegen weniger Probleme (Intervall zwischen 400 und 500 Punkten).

Hinweise zur Förderung

Die Testergebnisse zeigen Stärken und Schwächen auf. Kompetenzen und Begriffe, welche Intervallen zugeordnet sind, die tiefer als das Testergebnis sind, stellen für die Schülerinnen und Schüler in der Regel kein Problem dar. Kompetenzen und Begriffe, die Intervallen zugeordnet sind, in dem auch das Testergebnis liegt, sind bei den Schülerinnen und

Schülern vorhanden, sollten aber noch gefestigt werden. Kompetenzen und Begriffe, die Intervallen zugeordnet sind, die über dem Testergebnis liegen, sind bei den Schülerinnen und Schüler erst im Ansatz oder noch gar nicht vorhanden. Ihrer Vermittlung muss im Unterricht und beim selbstständigen Lernen der Schülerinnen und Schüler besondere Beachtung geschenkt werden.

6. Von den Testergebnissen zur Stellwerk-Skala

Eigenschaften der Normverteilung

Die Verteilung der Ergebnisse der Schülerinnen und Schüler in den Stellwerk-Tests nähern sich einer Normalverteilungskurve an. Die mittleren Werte (zwischen 400 und 600 Punkten) kommen häufig, die extrem hohen Werte (zwischen 700 und 800 Punkten) oder extrem tiefen Werte (zwischen 300 und 200 Punkten) kommen eher selten vor. Die Normalverteilungskurve hat die Form einer Glockenkurve. Ist die Glockenkurve flach und breit, dann ist die Streuung der Testergebnisse um den Mittelwert gross. Ist die Glockenkurve steil und schmal, dann ist die Streuung der Testergebnisse um den Mittelwert eher gering.

Die Normalverteilungskurven lassen sich anhand des Mittelwertes und der Streuung beschreiben. Der Mittelwert beziehungsweise das arithmetische Mittel entspricht der Summe aller Testergebnisse dividiert durch die Anzahl Testergebnisse. Die Streuung wird mit der Standardabweichung beziffert; ein Mass, das angibt, wie stark die Testergebnisse um den Mittelwert streuen. Die Standardabweichung wird aus den Abweichungen der einzelnen Testergebnisse vom Mittelwert berechnet. Sie entspricht der Wurzel aus dem Durchschnitt der quadrierten Abweichungen der Testergebnisse vom Mittelwert¹⁰.

$$SD = \sqrt{\frac{\sum_{j=1}^n (x_j - \bar{x})^2}{n}}$$

wobei SD = Standardabweichung (Streuung)

x_j = Testergebnis

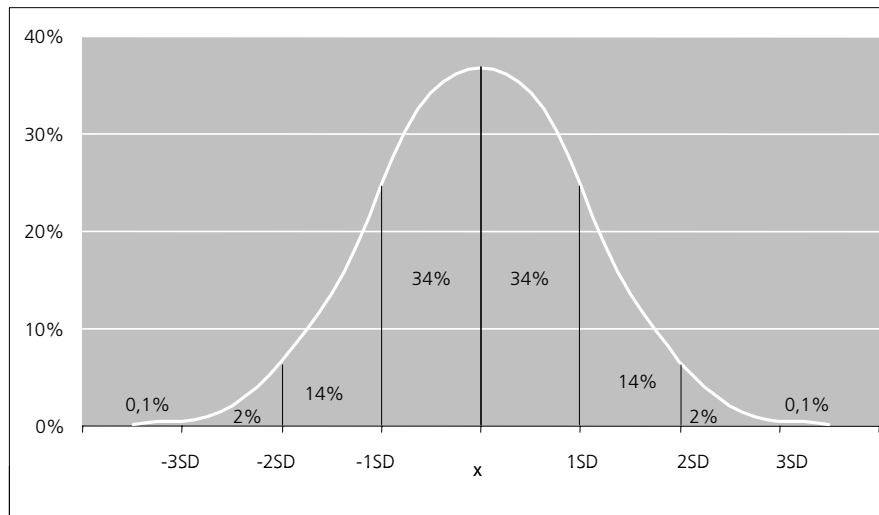
\bar{x} = Mittelwert

n = Stichprobenumfang

¹⁰ Weil der Mittelwert der Abweichungen bei einer Normalverteilung automatisch Null ergibt, werden die Abweichungen quadriert. Damit wird verhindert, dass die negativen Abweichungen durch die positiven kompensiert werden, was immer zu einer Streuung von $s = 0$ führen würde. Weil die Abweichungen quadriert wurden, wird zur Berechnung der Standardabweichung (Streuung) die Wurzel aus den durchschnittlichen Abweichungen gezogen.

Zu den Eigenschaften der Normalverteilungskurve gehört, dass rund 68 Prozent der Ergebnisse innerhalb von \pm einer Standardabweichung liegen. Rund 95 Prozent der Ergebnisse liegen innerhalb von \pm zwei Standardabweichungen und mehr als 99 Prozent der Ergebnisse liegen innerhalb von \pm drei Standardabweichungen (vgl. Abbildung 3).

Abbildung 3: Standardisierte Normalverteilung



Anmerkung: X = Mittelwert, SD = Standardabweichung (Streuung)

Die Normierung eines Mathematiktests bei Schülerinnen und Schülern der 8. Klasse könnte beispielsweise zu einem Mittelwert von 60 Prozent richtig gelöster Aufgaben und einer Standardabweichung von 10 Prozent führen. Das würde bedeuten, dass etwas mehr als zwei Drittel der Testergebnisse zwischen 50 und 70 Prozent richtig gelöster Aufgaben liegen, rund 95 Prozent zwischen 40 und 80 Prozent und nahezu alle Testergebnisse zwischen 30 und 90 Prozent.

Die Normierung eines Deutschtests bei den Schülerinnen und Schülern der 8. Klasse könnte beispielsweise zu einem Mittelwert von 40 Prozent richtig gelöster Aufgaben und einer Standardabweichung von 5 Prozent führen. Das würde bedeuten, dass etwas mehr als zwei Drittel der Testergebnisse zwischen 35 und 45 Prozent richtig gelöster Aufgaben liegen, rund 95 Prozent zwischen 30 und 50 Prozent und nahezu alle Testergebnisse zwischen 25 und 55 Prozent.

Testergebnisse als standardisierte z-Werte (Standardnormalverteilung)

Wie sind diese Unterschiede zwischen den Ergebnissen im Deutsch- und im Mathematiktest zu interpretieren? Damit Testergebnisse in Form der Abweichungen vom Mittelwert besser miteinander verglichen werden können, müssen sie zuerst an der Unterschiedlichkeit aller Werte der Population relativiert werden. Dies geschieht, indem die Abweichungen der individuellen Testergebnisse vom Mittelwert der Population abgezogen und durch die Standardabweichung der Population dividiert werden. Zu diesem Zweck werden die

sogenannten z-Werte als standardisierte Abweichungswerte nach folgender Formel berechnet¹¹:

$$z_i = \frac{x_i - \bar{x}}{s}$$

wobei z_i = Testergebnis der Person i als z-Wert

x_i = Testergebnis der Person i (beispielsweise als Anteil richtig gelöster Aufgaben)

\bar{x} = Mittelwert

SD = Standardabweichung (Streuung)

Die z-transformierte Verteilung wird auch Standardnormalverteilung genannt. Sie hat einen Mittelwert von $M = 0$ und eine Standardabweichung von $SD = 1$.

Wenn nun Schüler A im Mathematiktest 50 Prozent der Aufgaben richtig gelöst hat und Schüler B im Deutshtest ebenfalls 50 Prozent der Aufgaben richtig gelöst hat, erhalten wir durch die z-Transformation folgende Vergleichswerte:

$$z_A = \frac{50 - 60}{10} = -1$$

$$z_B = \frac{50 - 40}{5} = +2$$

Das Testergebnis von Schüler B im Deutshtest ist um drei Standardabweichungen besser als das Testergebnis von Schüler A im Mathematiktest. Nur 18 Prozent der Population erreichen schlechtere Ergebnisse im Mathematiktest als Schüler A. Nur 2 Prozent der Population erreichen bessere Ergebnisse im Deutshtest als Schüler B.

Testergebnisse auf der Stellwerk-Skala

Die Standardnormalverteilung wird durch den Mittelwert $M = 0$ und die Standardabweichung $S = 1$ definiert (vgl. Abbildung 3). Diese Darstellungsform hat den Nachteil, dass die Testergebnisse negativ sein können. Mittelwert und Standardabweichungen der standardisierten Normalverteilung lassen sich beliebig transformieren. Am bekanntesten sind vermutlich die Intelligenztestskalen, die auf einen Mittelwert von $M = 100$ Punkte und eine Standardabweichung von $S = 15$ Punkte normiert werden. Dies wird durch eine einfache Transformation erreicht, indem der z-Wert mit 15 multipliziert wird und anschliessend zum Ergebnis 100 Punkte addiert werden. Bei Stellwerk wurde eine andere Skala bevorzugt, was eine andere Transformation der z-Werte zur Folge. Der z-Wert wird mit 100 multipliziert und anschliessend werden zum Ergebnis 500 Punkte addiert.

$$\text{Testergebnis Stellwerk } T_{\text{Stellwerk}} = (100 \cdot z_{\text{Stellwerk}}) + 500$$

Die Ergebnisse liegen somit in der Regel zwischen 200 und 800 Punkten.

¹¹ Durch dieses Vorgehen wird die Fläche unter der Normalverteilungskurve in Einheiten der Standardabweichung unterteilt.