Raphael Zender et al. (Hrsg.): Die 18. Fachtagung Bildungstechnologien (DELFI),
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 307

Digitalisierung von Analyse- und Auswertungsstrukturen im Kontext schulischer Wettbewerbsszenarien

Florian Funke [1] und Sven Hofmann[2]

Abstract: Im Rahmen des Mathematik-Wettbewerbs Run For Numbers nehmen Schülerinnen und Schüler an Schulen in den Bundesländern Sachsen, Thüringen und Brandenburg halbjährlich teil, um ihre mathematischen Fähigkeiten und Fertigkeiten zu testen. Mit Hilfe des siebenminütigen Speed-Tests soll ein möglichst umfassendes Bild des aktuellen Lernstands erfasst werden. Im folgenden Artikel werden die umgesetzten Maßnahmen zur Digitalisierung der Auswertung und Analyse der Wettbewerbsdaten sowie dabei eingesetzte Kenngrößen, Maße und Normen beschrieben. Es wird ebenso dargestellt, inwieweit die Verständlichkeit der Werte für Lernende und Lehrende im schulischen Kontext sichergestellt wird und welche Darstellungsformen und Formulierungen eingesetzt werden.

Keywords: Assessment, Digitalisierung, Datenanalyse, Schule, Wettbewerb, Feedback, automatisierte Rückmeldung

1 Einleitung

Um fachübergreifende mathematische Schlüsselkompetenzen zu trainieren und eine Rückmeldung zum aktuellen Lernstand zu geben, wurde der Wettbewerb Run For Numbers entwickelt. Damit dieser eine hohe Passung in die schulischen Rahmenbedingungen aufweist, wurde er als siebenminütiger Speed-Test konzipiert. Eine Einbettung des Wettbewerbs als Teil einer Unterrichtseinheit ist damit sehr gut möglich. Zur Qualitätssicherung und iterativen Verbesserung werden Beurteilung und Bewertung der Lösungen, sowie die Auswertung und Analyse zentral von der Wettbewerbsorganisation durchgeführt. Für eine Anpassung an individuelle Voraussetzungen wird die Durchführung digital und papierbasiert angeboten [Fu19a]. Durch die laufend veränderlichen Rahmenbedingungen für die Durchführung, beispielsweise in der Ausstattung der Schulen, wurden Möglichkeiten untersucht, die Analyse und Auswertung der Wettbewerbsdaten unabhängig von der Art und Weise der Durchführung zu gestalten. Damit wird sichergestellt, dass bei Veränderungen der Rahmenbedingungen des Wettbewerbs keine umfassende Rekonstruktion der Analysestrukturen erfolgen muss. Im Folgenden werden die implementierten Funktionen und die verwendeten Methoden erläutert sowie die daraus generierten Feedbacks für Lernende und Lehrende beschrieben.

[1] Universität Leipzig, Didaktik der Informatik, Augustusplatz 10, 04109 Leipzig,
florian.funke@informatik.uni-leipzig.de, https://orcid.org/0000-0003-4043-894X
[2] Universität Leipzig, Didaktik der Informatik, Augustusplatz 10, 04109 Leipzig,
sven.hofmann@informatik.uni-leipzig.de

2 Datenerhebung und Datenverarbeitung

Die Konzeption der Analysewerkzeuge sollte mehrere veränderliche Rahmenbedingungen des Wettbewerbs berücksichtigen. Zum einen findet der Wettbewerb nach wie vor auf Grund unterschiedlicher Ausstattungen und Voraussetzungen an Schulen papierbasiert und digital statt. Parallel dazu steigen die Teilnehmerzahlen kontinuierlich, auf zuletzt über 5000 teilnehmende Schülerinnen und Schüler, an [Fu19a]. Andererseits sind neue Wege der Rückmeldung auf verschiedenen Ebenen eingebunden worden. So sind nun Möglichkeiten des individualisierten Feedbacks für Teilnehmende und Lehrende etabliert. Mit Hilfe der neu geschaffenen Strukturen wird darüber hinaus die Qualitätssicherung des Wettbewerbs vorgenommen. Der Ablauf der Datenerhebung ist in Abb. 1 dargestellt.

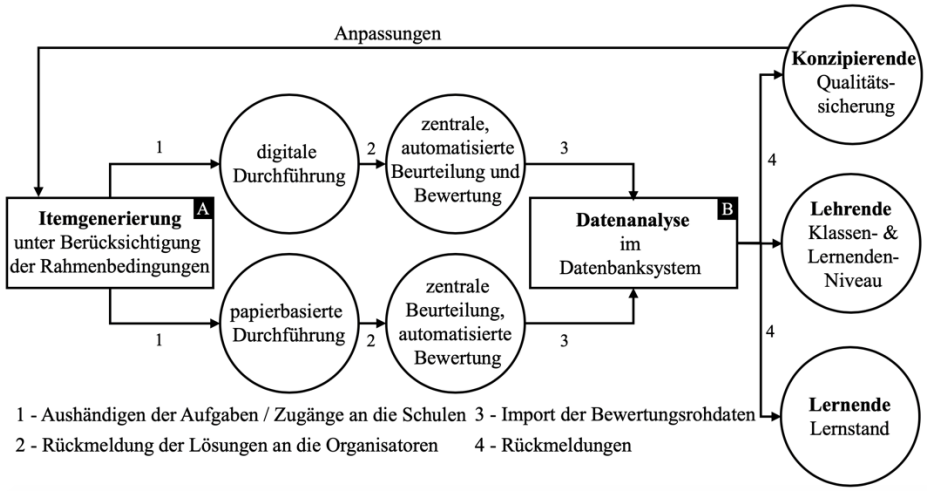

Abb. 1: Ablauf der Datenerhebung im Rahmen des Run For Numbers

3 Datenanalyse

Um die auf verschiedenen Wegen erhobenen Daten zentral analysieren zu können, wurde ein Datenbanksystem entwickelt, in dem sich die Rohdaten mittels Import-Funktion, aber auch händisch einpflegen lassen. Mit Hilfe von Einweg-Hash-Funktionen wird ermöglicht, dass die Daten eines Teilnehmers ihm nicht mehr zuzuordnen sind. Da die Berechnung aber über die Wettbewerbe hinweg aus den gleichen Metadaten erfolgt, ist weiterhin die Zuordnung von verschiedenen Teilnahmen am Wettbewerb zu einem Teilnehmer möglich. Veränderungen des individuellen Lernstands sind damit in zukünftig entwickelten Darstellungsformen verwendbar und können rückgemeldet werden. Die innerhalb eines Wettbewerbs erhobenen Daten werden dabei auf Ebene der Items, der Teilnehmer und der Klasse untersucht.

3.1 Itemanalyse

Die Itemanalyse leistet einen Beitrag hinsichtlich der Untersuchung von Leistungen in Korrelation mit der Gesamtstichprobe sowie bei der Erhöhung der Validität und damit in der Qualitätssicherung. Die hohe Validität eines Itemsatzes impliziert, dass Personen mit starker Merkmalsausprägung die angegebenen Aufgaben häufiger richtig beantworten, als Personen, die eine schwache Merkmalsausprägung haben [Li89]. Die Itemanalyse kann dabei zur kontinuierlichen Erhöhung der Validität über mehrere Iterationen beitragen. Die Validität bestimmt sich über die Objektivität, Reliabilität, Schwierigkeit und Trennschärfe [Li89]. Bei der Bestimmung dieser Größen erfolgt eine Orientierung an der klassischen Testtheorie, da sich die untersuchte mathematische Kompetenz aus verschiedenen Dimensionen zusammensetzt. Eine Modellierung konformer Items im Rahmen der Item-Response-Theorie ist mit dieser Annahme nicht möglich [Ro04].

Die Prüfung der Objektivität erfolgt in den Dimensionen Durchführung, Auswertung und Interpretation. Die Sicherstellung erfolgt mit Hilfe von standardisierten Abläufen und Formulierungen, die gleichwertige Bedingungen für alle Wettbewerbsteilnehmer garantieren sollen. Insbesondere im Rahmen der zentralen, teilautomatisierten Beurteilung und Bewertung kann die Objektivität sehr gut sichergestellt werden. Dabei hat die Formalisierung der Aufgabenstellungen hohe Relevanz, um mögliche Alternativlösungen zu eliminieren und damit die automatisierte Beurteilung und Bewertung zu ermöglichen. Auf Basis der stattfindenden Datenanalyse wurde die Objektivität kontinuierlich von Wettbewerb zu Wettbewerb verbessert. Dabei erfolgt die Formalisierung der verwendeten Items in mehreren Schleifen. Der Ablauf einer Iteration ist in Abb. 1 dargestellt. Die im Folgenden beschriebenen Kenngrößen der Itemanalyse bieten wertvolle Indikatoren, um problematische Items zu identifizieren, anzupassen oder zu eliminieren.

Für die Bestimmung der Schwierigkeit eines Items wird der Schwierigkeitsindex herangezogen. Dieser gibt den Anteil richtiger Antworten eines Items an [Li89]. Der Schwierigkeitsindex findet unter anderem Anwendung in der Analyse einer Klasse im Vergleich zur Gesamtstichprobe. Lehrende können schnell vergleichen, wie ein Item in der Klasse im Vergleich zum gesamten Teilnehmerfeld beantwortet wurde. Daraus können Rückschlüsse über den Lernstand der Klasse im untersuchten mathematischen Teilgebiet gezogen und Aufgabentypen, die zu leicht oder zu schwer sind, identifiziert werden.

Die Trennschärfe beschreibt die Korrelation der Beantwortung eines Items mit der Gesamtpunktzahl des Teilnehmenden [Fi04]. Im Rahmen der Test- und Itemkonstruktion wird auf Basis vorheriger Wettbewerbe versucht, Items mit einer Trennschärfe im Bereich .5 bis .9 zu generieren. Dies erfolgt, da Items mit zu hoher Trennschärfe (>.9) wenig neue Informationen über die Merkmalsausprägung liefern und Items mit geringer Trennschärfe (<.5) andere Eigenschaften, als die Anvisierten bestimmen [Ro04]. Damit wird eine gute Differenzierung zwischen den Teilnehmenden ermöglicht. Dies gelingt noch nicht umfänglich. Im Herbst 2019 waren lediglich jedes Zehnte der Items zwischen .5 und .9, aber zwei von drei Items in der Umgebung des anvisierten Bereichs (.4 bis .95) [Fu19b].

Für die Bestimmung der Reliabilität wird das Kuder-Richardson-Alpha verwendet. Dieses gibt an, inwieweit die Items innerhalb einer Testung Gleiches messen und die Äquivalenz sowie Stabilität von Messungen gewährleistet ist [Li89]. Mit Hilfe des berechneten Maßes ist es dementsprechend möglich die Qualität des Itemsets einzuschätzen. Dabei werden Werte zwischen .65 und .95 für Testungen als gut angenommen [GM02]. Im Herbst 2019 wurden Werte zwischen .76 und .84 erreicht [Fu19b].

3.2 Teilnehmeranalyse

Im Rahmen der Analyse der Teilnehmer soll eine zweistufige Verortung der Leistung eines Teilnehmers im gesamten Teilnehmerfeld vollzogen werden. Es kann in der Auswertung ein Vergleich mit zurückliegenden Leistungen (intraindividuelle Norm), und mit Leistungen anderer Gruppen im aktuellen Wettbewerb (interindividuelle Norm) erfolgen. Vergleichsgruppen sind hierbei Klassen, Schulen und Schulformen.

Im ersten Schritt wird der Prozentrang jedes Teilnehmenden ermittelt. Dieser gibt an, wie viel Prozent aller Teilnehmer die gleiche oder eine geringere Punktzahl erreicht haben [Li89]. Der Prozentrang wird daher zur Verortung der individuellen Leistung eines Teilnehmenden in der Stichprobe eingesetzt. Der entstandene Wert wird an Lernende und Lehrende gleichermaßen rückgemeldet. Mit den Werten der Lernenden können der aktuelle Lernstand sowie Tendenzen in der Entwicklung abgelesen werden. Der Prozentrang ist mit dem Charakter einer Stichprobe versehen, da er nicht normiert ist, auch wenn die Verteilung der Teilnehmer einer Normalverteilung in der Regel ähnelt [Fu19b].

Zur Normierung wird aus diesem Grund der T-Wert als weiteres Maß aus dem Prozentrang berechnet. Die T-Wert-Skala eignet sich in diesem Fall, da sie immer dann Anwendung findet, wenn die Messwerte keine Normalverteilung liefern, aber eine solche für das untersuchte Merkmal angenommen wird [HL11]. Damit wird mit Hilfe des T-Werts, ähnlich dem Prozentrang, die Position eines Teilnehmers innerhalb der gesamten Messung widergespiegelt, gleichzeitig aber der Stichprobencharakter eliminiert.

Aus den beschriebenen Kenngrößen werden über die individuelle Rückmeldung hinaus Klassenwerte bestimmt. Dabei eignet sich der Mittelwert der T-Werte besonders, um die Klasse im Vergleich zur Gesamtstichprobe zu betrachten. Mit Hilfe des Prozentrangs wird dargestellt, wie viel Prozent der Teilnehmer einer Klasse im Schnitt überboten wurden.

4 Aufbereitung für Teilnehmende und Lehrende

Für Empfänger, die im Umgang mit den berechneten Werten und ihren Bedeutungen keine Erfahrung besitzen, müssen diese interpretierbar gemacht werden. Auch stellt sich die Frage, welche Werte in welcher Form rückgemeldet werden. Die Auswertung erfolgt dabei in zwei Parallelen für die Lernenden und Lehrenden (vgl. Abb. 2).

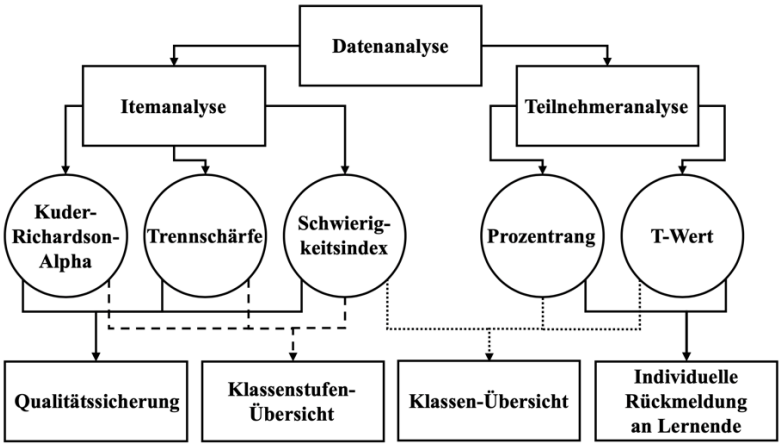

Abb. 2: Berechnung und Verwendung der Kenngrößen

Für die Teilnehmenden und ihre Orientierung an interindividuellen Normen stellen der Prozentrang und der T-Wert zentrale Rückmeldungen dar. Dabei lässt sich der Prozentrang gut in eine leichter interpretierbare Kenngröße überführen. Den Schülerinnen und Schülern wird die Aussage: „Du gehörst zu den besten x% deiner Klassenstufe." rückgemeldet. Diese ist für sie besser verständlich und gut interpretierbar.

Aus dem T-Wert lässt sich ein Ausschnitt des Teilnehmerfeldes ermitteln, in dem sich die Teilnehmenden bewegen. Anhand festgelegter Bereiche, in denen sich der T-Wert eines Teilnehmers befindet, werden verschiedene sprachliche Rückmeldungen gegeben. Für Teilnehmende mit einem T-Wert im Bereich 60 bis 70 wird beispielsweise „Du hast herausragende Leistungen erbracht" rückgemeldet.

Für die Lehrenden soll ein umfassender statistischer Überblick zum Wettbewerb, zur Klasse und für die einzelnen Lernenden möglich sein. Daher werden zunächst die Daten der Itemanalyse und statistische Daten, wie die im Mittel erreichten Punkte und deren Verteilung in einer Klassenstufenübersicht aufbereitet. Hierbei werden Items mit hohem und niedrigem Schwierigkeitsindex farblich hervorgehoben, so dass beispielsweise besonders schwere Aufgaben in der Klasse zur Diskussion herausgehoben werden können.

Darüber hinaus erhalten die Lehrenden eine Übersicht, wie die Klasse Aufgaben im Mittel beantwortet hat. Hierbei werden analog Abweichungen von der Gesamtbeantwortung und damit dem Schwierigkeitsindex hervorgehoben. Hier werden klassenspezifische Exzellenz- und Potenzialgebiete dargestellt. Abschließend wird eine Übersicht mit den Rohwerten, T-Werten und Prozenträngen sowie den zugehörigen generierten sprachlichen Formulierungen zur Verfügung gestellt, so dass auch die individuellen Werte der Lernenden von den Lehrenden untersucht werden können. Die Interpretierbarkeit der Ergebnisse wird dabei durch eine Kurzhandreichung sichergestellt, die die Hintergründe der einzelnen Werte darstellt. Alle beschriebenen Übersichten werden automatisch generiert und in eine gut verteilbare Form, beispielsweise Serienbriefe, gebracht.

5 Fazit und Ausblick

Die für den Wettbewerb Run For Numbers umgesetzten zentralisierten, digitalen Analysestrukturen konnten etabliert werden. Sie haben sich in den Wettbewerben als hilfreich bei der Generierung von individualisiertem Feedback in verschiedenen Dimensionen gezeigt. Durch sie konnten bei kontinuierlich steigenden Teilnehmerzahlen umfassendere und hilfreiche Auswertungen für Lernende und Lehrende generiert werden. Die Rückmeldungen der teilnehmenden Schulen zeigen auf, dass durch die regelmäßige Durchführung, und dass damit einhergehende Feedback an die Lernenden, gezielt individuelle Förderungen angestoßen werden können. Die eingesetzten Erläuterungen und Verbalisierungen erzeugen eine Verständlichkeit der statistischen Werte und machen sie erst dadurch insbesondere für die Lernenden nutzbar.

Im Rahmen der Weiterentwicklung sollen gezielt auch wettbewerbsübergreifende Daten ausgewertet und in einer intuitiven Form dargestellt werden, damit die Lernenden und Lehrenden nicht nur eine Rückmeldung zum aktuellen Lernstand, sondern auch zur Entwicklung im Vergleich zu vorangegangen Wettbewerben erhalten. Hierbei können die bereits verwendeten statistischen Werte eingebunden und visualisiert werden. Darüber hinaus werden Konzepte zur Abbildung der Erfüllung kriterialer Normen entwickelt und erprobt.

Literaturverzeichnis

[Fi04] Fisseni, H.: Lehrbuch der psychologischen Diagnostik. Göttingen: Hogrefe, S. 36ff, 2004.

[Fu19a] Funke, F.: Entwicklung und Evaluation der iterativen Digitalisierung kompetitiver Assessment-Szenarien dargestellt am Beispiel des Wettbewerbs Run For Numbers, 2019.

[Fu19b] Funke, F.: Ergebnisse Herbst 2019. https://runfornumbers.de/ergebnisse-herbst-2019/, Stand: 26.03.2020

[GM02] George, D.; Mallery, P.: SPSS for Windows Step by Step. Needham Heights, MA: Pearson Higher Education, S.231, 2002.

[HL11] Hesse, I.; Latzko, B.: Diagnostik für Lehrkräfte. Opladen: Budrich, S. 75, 77, 2011.

[Li89] Lienert, G.: Testaufbau und Testanalyse. München: Psychologie-Verl.-Union, S. 38, 40, 70, 226, 331ff., 1989.

[RK39] Richardson, M. W.; Kuder, G. F.: The Calculations of Test Reliability Coefficients Based on the Method of Rational Equivalence. J. Educ. Psychol, S. 30, 681, 1939.

[Ro04] Rost, J.: Lehrbuch Testtheorie - Testkonstruktion. Bern: Huber. S. 218, 2004.