Wie vergleicht man den Anspruch mathematischer Prüfungen?

Die *A levels* in England, Wales und Nordirland

Ian Jones, Chris Wheadon, Sara Humphries und Matthew Inglis

Die *Certificate of Education Advanced Level*-Prüfungen (kurz: *A level*) in Mathematik entsprechen im Sekundarschulsystem von England, Wales und Nordirland ungefähr den deutschen Abiturklausuren in Mathematik. Wie im Falle der Abiturklausuren in Deutschland gibt es auch bei diesen Prüfungen eine öffentlich diskutierte Besorgnis, dass ihre Ansprüche über die Jahre sinken, und dass daher die Studienanfänger ohne die notwendigen Vorkenntnisse und Fertigkeiten an die Universitäten kommen. Allerdings ist ein Vergleich der Ansprüche von Prüfungen aus verschiedenen Jahrgängen und allgemein von Prüfungen, die nicht aus demselben kurrikularen Kontext stammen, eine methodisch schwierige Aufgabe; Behauptungen über sinkende Standards der *A level*-Mathematikprüfungen stützen sich daher in der Regel auf Einzelfallbeispiele sowie persönliche Überzeugungen und weniger auf empirische und belastbare Forschungsergebnisse.

In einer jüngst im *British Educational Research Journal* veröffentlichten Studie beschäftigen wir uns mit dem methodisch schwierigen Problem des Vergleichs der Ansprüche von *A level*-Mathematikprüfungen über fünf Jahrzehnte. Verwendet wurde eine neuartige Methode auf der Grundlage des *Paarvergleichs*, von der wir denken, dass sie eine Lösung des genannten methodischen Problems darstellt.

Bevor wir im Folgenden diese Methode vorstellen, möchten wir zunächst den deutschen Leserinnen und Lesern den Hintergrund und Zusammenhang der *A level*-Mathematikprüfungen erläutern.

Die A levels

In Bezug auf die Mathematikausbildung sind England, Wales und Nordirland ungewöhnlich, da Schülerinnen und Schüler im Alter von 16 Jahren Mathematik vollständig abwählen und damit ihre Mathematikausbildung beenden können. In ihren letzten zwei Schuljahren – also im Alter von 16 bis 18 Jahren – wählen Schülerinnen und Schüler üblicherweise lediglich drei oder vier Fächer aus, von denen ein oder zwei Fächer (s. u.) Mathematik sein können, aber nicht müssen. Die Kurse dieser letzten zwei Schuljahre werden als *A level*-Kurse bezeichnet und die Noten in den zugehörigen *A level*-Prüfungen werden von den Universitäten als Auswahlkriterium verwendet. Traditionell gab es Noten von A (die beste Note) bis E; 2008 wurde zusätzlich die neue Note A* eingeführt, um einer Noteninflation entgegenzuwirken. Schülerinnen und Schüler, die den *A level*-Kurs in Mathematik belegen, lernen Differential- und Integralrechnung, Trigonometrie, Koordinatengeometrie, Algebra und Mechanik. Der Kurs behandelt keine Beweise und Beweistechniken, komplexe Zahlen oder lineare Algebra; diese werden in einem zweiten *A level*-Kurs mit dem Namen „weitere Mathematik" (*further mathematics*) unterrichtet. Ein Mathematikstudiengang einer hochangesehenen britischen Universität erwartet im Auswahlverfahren von Bewerberinnen und Bewerbern in der Regel zwei oder drei *A levels* mit der Note A*, darunter sowohl Mathematik als auch weitere Mathematik; weniger angesehene Universitäten geben sich mit der Note B zufrieden und verlangen ggf. nicht den Kurs in weiterer Mathematik.

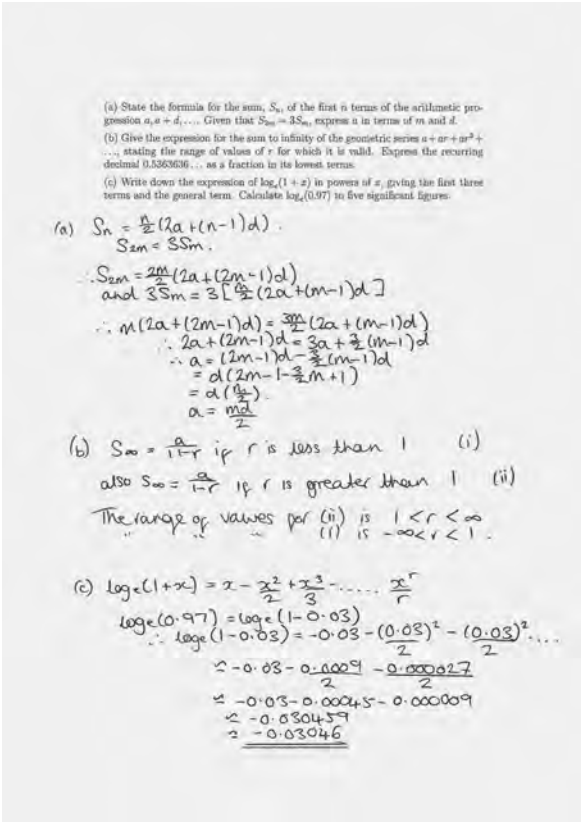
A level-Kurse und -Prüfungen in Mathematik gibt es seit den 1950er Jahren. Aber sind ihre Standards über die Jahrzehnte gleichgeblieben? Mit anderen Worten: Sind die mathematischen Fertigkeiten einer Bewerberin oder eines Bewerbers mit einer Note A aus den 1950er Jahren vergleichbar mit denen einer heutigen Bewerberin oder eines heutigen Bewerbers mit der Note A? Diese Frage ist nicht einfach zu beantworten.

Ansätze zum Vergleich von Prüfungsstandards

Ein naheliegender Ansatz ist, Prüflinge aus verschiedenen Jahren mittels eines einheitlichen Tests zu prüfen: Wenn Personen mit identischen Testergebnissen im einheitlichen Test unterschiedliche Noten in den *A level*-Prüfungen haben, so liefert dies einen Hinweis, dass sich die Standards verändert haben [2, 5]. Allerdings verändern sich nicht nur die Prüfungen, sondern auch der Inhalt der *A level*-Kurse. Die Methode des einheitlichen Tests überprüft daher weniger die Standards verschiedener Jahre, sondern eher, ob die *A level*-Kurse in den verschiedenen Jahren eine gute Vorbereitung auf diesen gewählten einheitlichen Test sind.

Eine weiterer Ansatz wäre, heutigen Absolventen die Prüfungen vergangener Jahre zu geben; auch dieser Ansatz ist problematisch, da man den Kurs als Prüfungsvorbereitung – sowohl inhaltlich als auch die Vorbereitung auf den spezifischen Stil der Prüfungsfragen – nicht von der Prüfung trennen kann. Schlechtere Prüfungsleistun-

 DOI 10.1515/dmvm-2016-0039



(a) State the formula for the sum, S_n, of the first n terms of the arithmetic progression $a, a+d, \ldots$. Given that $S_{2m} = 3S_m$, express a in terms of m and d.

(b) Give the expression for the sum to infinity of the geometric series $a + ar + ar^2 + \ldots$, stating the range of values of r for which it is valid. Express the recurring decimal $0.5363636\ldots$ as a fraction in its lowest terms.

(c) Write down the expression of $\log_e(1+x)$ in powers of x, giving the first three terms and the general term. Calculate $\log_e(0.97)$ to five significant figures.

(a) $S_n = \frac{n}{2}(2a+(n-1)d)$.
$S_{2m} = 3S_m$.
$\therefore S_{2m} = \frac{2m}{2}(2a+(2m-1)d)$
and $3S_m = 3\left[\frac{m}{2}(2a+(m-1)d\right]$
$\therefore m(2a+(2m-1)d) = \frac{3m}{2}(2a+(m-1)d)$
$\therefore 2a+(2m-1)d = 3a+\frac{3}{2}(m-1)d$
$\therefore a = (2m-1)d - \frac{3}{2}(m-1)d$
$= d(2m-1-\frac{3}{2}m+1)$
$= d(\frac{m}{2})$.
$a = \frac{md}{2}$

(b) $S_\infty = \frac{a}{1-r}$ if r is less than 1 (i)
also $S_\infty = \frac{a}{1-r}$ if r is greater than 1 (ii)
The range of values for (ii) is $1 < r < \infty$
" " " (i) is $-\infty < r < 1$.

(c) $\log_e(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots \frac{x^r}{r}$
$\log_e(0.97) = \log_e(1-0.03)$
$\therefore \log_e(1-0.03) = -0.03 - \frac{(0.03)^2}{2} - \frac{(0.03)^3}{2} \ldots$
$\simeq -0.03 - 0.0009 - 0.000027$
$\simeq -0.03 - 0.00045 - 0.000009$
$\simeq -0.030459$
$\simeq -0.03046$

(a) Express $\sin x - 3\cos x$ in the form $R\sin(x-\alpha)$, where $R > 0$ and $0° < \alpha < 90°$, giving your value of α to the nearest $0.1°$.

(b) Hence find the values of x in the interval $0° < x < 360°$ for which

$$\sin x - 3\cos x + 2 = 0$$

giving your values of x to the nearest degree.

a. $\sin x - 3\cos x \quad R\sin(x-\alpha)$
$\sin(A-B) = \sin A\cos B - \cos A\sin B$
$R\cos\alpha = 1$, $R\sin\alpha = 3$ $\Big\}$ $\sqrt{3^2+1^2} = \sqrt{10}$ $R = \sqrt{10}$
$\tan\alpha = \frac{3}{1}$ $\tan\alpha = 3$ $\alpha = 71.56 = 71.57$ 2dp
$\sin x - 3\cos x = \sqrt{10}\sin(x-71.57)$

b. $\sin x - 3\cos x + 2 = 0$
$\sin x - 3\cos x = -2$
$\sqrt{10}\sin(x-71.57) = -2$
$\sin(x-71.57) = -\frac{2}{\sqrt{10}}$ or $\frac{\sqrt{10}}{5}$
$x - 71.57 = 39.23, 219.23,$
$x = 111°, 291°$

Abbildung 1. *Zwei schriftliche Prüfungsleistungen:* Which candidate is the better mathematician?

gen in einer Prüfung, auf die man nicht vorbereitet wurde, sind nicht notwendigerweise ein aussagekräftiges Argument für geringere mathematische Fertigkeiten.

Die Methode des Paarvergleichs

Unser Ansatz verwendet die Methode des *Paarvergleichs*. Diese Methode basiert auf der Erkenntnis, dass menschliche Urteile beim direkten Vergleich zweier Objekte – in Bezug auf viele verschiedene Reize – deutlich genauer sind als solche, bei denen ein Objekt isoliert betrachtet wird [6]. Zum Beispiel ist es einfacher zu entscheiden, in welchem von zwei Räumen es wärmer ist, als die Temperatur in einem der Räume zu schätzen. Im frühen zwanzigsten Jahrhundert begründete der amerikanische Psychometriker Louis Leon Thurstone (1887–1955) die Theorie des Paarvergleichs; inzwischen gibt es gute psychophysische Modelle für die Mechanismen, die hinter diesen binären Urteilen stehen.

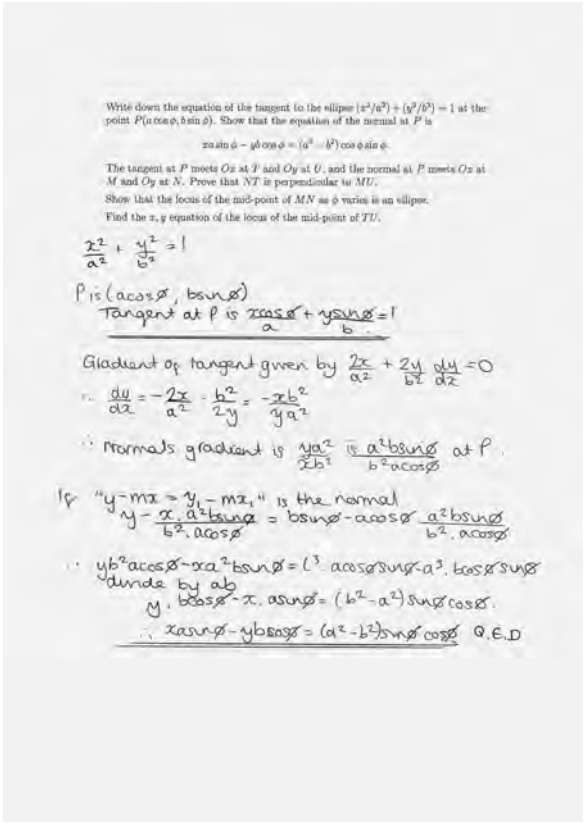
In unserem Kontext, erhalten Evaluatoren in einem solchen Paarvergleich schriftliche Prüfungsleistungen zweier Prüflinge und werden aufgefordert, die bessere Leistung zu identifizieren. Das Konstrukt „bessere Leistung" wird dabei operationalisiert durch die abstrakte Frage: *Which student do you think is the better mathematician*? Ein Evaluator sieht also zwei Prüfungsleistungen vor sich, eine auf der linken und eine auf der rechten Seite und beantwortet die Frage entweder mit *links* oder *rechts*.

Die Ergebnisse einer Vielzahl solcher Paarvergleiche können dann statistisch mit dem sogenannten Bradley–Terry-Modell analysiert werden ([1]) und liefern eine relative Parameterschätzung der Qualität der schriftlichen Prüfungsleistung. Wir nennen diesen Parameter im folgenden kurz „Leistung". Wir untersuchen die Übereinstimmung der Urteile zwischen verschiedenen Evaluatoren, um zu bestimmen, ob die Evaluatoren dasselbe Konstrukt bewertet haben. Mit anderen Worten: Dieser Prozess erlaubt es uns, zu bewerten, in welchem Maß die Beurteilung des Paarvergleichs gemäß dem Kriterium *better mathematician* zwischen verschiedenen Evaluatoren übereinstimmt.

Unsere Studie

In unserer Arbeit haben wir 66 schriftliche Prüfungsleistungen aus *A level*-Prüfungen der Jahre 1964, 1968, 1996 und 2012 aus den Archiven des *Office of Qualifications and Examinations Regulation* (Ofqual) erhalten. Es wäre vorteilhaft gewesen, Daten aus mehr Jahrgängen zu haben, aber leider ist die Abdeckung des Ofqual-Archivs hierfür nicht ausreichend; insbesondere enthält es keine schriftlichen Prüfungsleistungen aus den 1970er, 1980er und 2000er Jahren.

Die 66 Prüfungsleistungen waren mit den Noten A, B und E bewertet worden, und wir haben nur Prüfungsleistun-



Write down the equation of the tangent to the ellipse $(x^2/a^2)+(y^2/b^2)=1$ at the point $P(a\cos\phi, b\sin\phi)$. Show that the equation of the normal at P is

$$ax\sin\phi - by\cos\phi = (a^2-b^2)\cos\phi\sin\phi.$$

The tangent at P meets Ox at T and Oy at U, and the normal at P meets Ox at M and Oy at N. Prove that NT is perpendicular to MU.

Show that the locus of the mid-point of MN as ϕ varies is an ellipse.

Find the x, y equation of the locus of the mid-point of TU.

$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$

P is $(a\cos\phi, b\sin\phi)$

Tangent at P is $\frac{x\cos\phi}{a} + \frac{y\sin\phi}{b} = 1$

Gradient of tangent given by $\frac{2x}{a^2} + \frac{2y}{b^2}\frac{dy}{dx} = 0$

$\therefore \frac{dy}{dx} = -\frac{2x}{a^2}\cdot\frac{b^2}{2y} = -\frac{xb^2}{ya^2}$

$\therefore$ Normals gradient is $\frac{ya^2}{xb^2} = \frac{a^2b\sin\phi}{b^2a\cos\phi}$ at P.

If "$y-mx = y_1 - mx_1$" is the normal

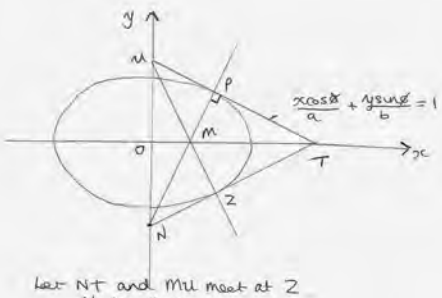
$y - \frac{x.a^2b\sin\phi}{b^2.a\cos\phi} = b\sin\phi - a\cos\phi\,\frac{a^2b\sin\phi}{b^2.a\cos\phi}$

$\therefore yb^2a\cos\phi - xa^2b\sin\phi = b^3a\cos\phi\sin\phi - a^3.b\cos\phi\sin\phi$

divide by ab

$y.b\cos\phi - x.a\sin\phi = (b^2-a^2)\sin\phi\cos\phi.$

$\therefore xa\sin\phi - yb\cos\phi = (a^2-b^2)\sin\phi\cos\phi$ Q.E.D



Let NT and MU meet at Z

At N, $x=0$

$\therefore -ybc\cos\phi = (a^2-b^2)a\sin\phi\cos\phi$

$\therefore y = \frac{(b^2-a^2)a\sin\phi}{b}$

At T, $y=0 \therefore x = \frac{a}{\cos\phi}$

$\therefore$ Gradient of TN is $\frac{0 - \frac{a\sin\phi}{b}(b^2-a^2)}{\frac{a}{\cos\phi} - 0}$

$= \frac{-\sin\phi(b^2-a^2)\cos\phi}{ab}$

At M, $y=0 \therefore x = \frac{(a^2-b^2)\sin\phi\cos\phi}{a\sin\phi}$

At U $x=0, \therefore y = \frac{b}{\sin\phi}$

$\therefore$ Grad of MU is $\frac{\frac{b}{\sin\phi}}{0-(a^2-b^2)\cos\phi}$

$= \frac{b^2a}{(b^2-a^2)\sin\phi\cos\phi}$

Grad MU × Grad TN $= \frac{ab(b^2-a^2)\sin\phi\cos\phi}{(b^2-a^2)\sin\phi\cos\phi ab}$

$= -1$

Abbildung 2. *Oben und folgende Seite: Aus einer weiteren Prüfungsleistung aus dem Jahr 1968*

gen berücksichtigt, die nicht dicht an den Benotungsgrenzen lagen. Die 66 Prüfungsleistungen wurden in 546 einzelne Fragen aufgeteilt, welche die Grundlage für unseren Paarvergleich waren. Da wir befürchteten, dass Typografie und Handschrift das Alter der Dokumente verraten könnten, setzten wir alle Fragen neu und schrieben die 546 Antworten in einer einheitlichen Handschrift ab.

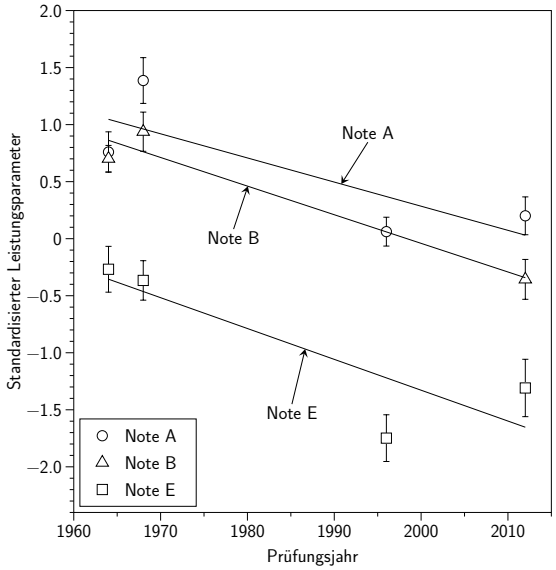
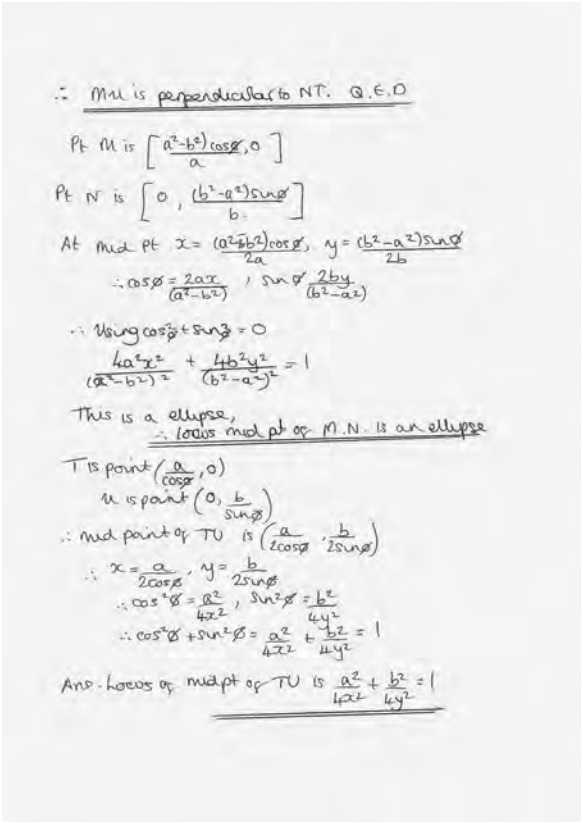


Abbildung 3. *Mittlerer standardisierter Parameterschätzer für die schriftlichen Prüfungsleistungen für jede Note und jedes Jahr. Die Fehlerbalken zeigen* ± 1 *Standardabweichung.*

Die zwanzig Evaluatoren waren Doktorandinnen und Doktoranden der Mathematik. Sie wurden nicht über den Zweck der Studie informiert und die Ergebnisse einer Befragung der Evaluatoren nach der Studie zeigen, dass keiner der Evaluatoren geahnt hat, dass wir uns mit dem Vergleich von Prüfungsstandards in verschiedenen Jahren befassen. Die Evaluatoren verwendeten ein Onlinesystem für Paarvergleich (www.nomoremarking.com). Bei jedem Vergleich sah ein Evaluator zwei schriftliche Leistungen (ein Beispiel findet sich in Abbildung 1) und musste entscheiden, welcher der beiden Prüflinge *the better mathematician* ist.

Insgesamt erhielten wir 5000 solcher Urteile, die wir mit dem Bradley–Terry 2-Paket in R ([1, 3]) anpassten, und wiesen so jeder Antwort eine standardisierte Parameterschätzung der Leistung zu. Wir bewerteten auch die interne Konsistenz des Modellierungsprozesses mit mehreren Methoden: Jede dieser Bewertungen ergab, dass unsere Methode zuverlässig ist (die statistischen Details finden sich in [4]). Der Mittelwert der Parameterschätzungen aller vom Prüfling bearbeiteten Fragen wurde als Maß der *Leistung* dieser Person verwendet.



Die Mittelwerte der Leistung für die einzelnen Noten in jedem der untersuchten Jahre finden sich in Abbildung 3. Wir führten eine multiple Regressionsanalyse mit der Leistung als abhängiger Variable sowie Note (A=5, B=4 und E=1) und Jahr als erklärender Variable durch: Gemäß dem Modell erklären Note und Jahr 74,8 % der Varianz der abhängigen Variable. Sowohl Note ($\beta = 0{,}69$, $p < 0{,}001$) als auch Jahr ($\beta = -0{,}51$, $p < 0{,}001$) waren signifikante Einflussgrößen. Wie man in Abbildung 3 erkennen kann, findet der Abfall der Standards im wesentlichen zwischen den sechziger und neunziger Jahren statt. Die Leistung eines Prüflings mit der Note B in den neunziger Jahren entspricht grob der Leistung eines Prüflings mit der Note E in den sechziger Jahren. Im Gegensatz zu Behauptungen der derzeitigen britischen Regierung findet sich kein Beleg für einen Abfall der Standards zwischen den neunziger Jahren und heute.

Insgesamt funktionierte der Paarvergleich in unserer Studie sehr gut. Die Evaluatoren hatten keine Schwierigkeiten mit unserem abstrakten Kriterium *better mathematician* und wir konnten die Daten an das Bradley–Terry-Modell ohne Probleme anpassen. Wir denken daher, dass Paarvergleiche eine sinnvolle Methode darstellen, um Ausbildungsstandards über verschiedene Kurrikula zu vergleichen. Für interessierte Leserinnen oder Leser, die selbst einige der schriftlichen *A level*-Prüfungsleistungen vergleichen wollen, stellen wir unsere Webseite unter http://tinyurl.com/JudgeMaths zur Verfügung. Allgemein steht das No More Marking-Instrument zur Verwendung bei Paarvergleichen von schriftlichen Prüfungsleistungen für nichtkommerzielle Verwendung unter http://www.nomoremarking.com zur Verfügung. Man kann diese Methode natürlich auch bei der Benotung von Studierenden im üblichen Lehrzusammenhang anwenden.

Literatur

[1] Bradley, R. and Terry, M. (1952). Rank analysis of incomplete block designs the method of paired comparisons. *Biometrika*, 39(3–4):324–345.

[2] Coe, R. (2007). Changes in standards at gcse and a level: Evidence from alis and yellis.

[3] Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical Software*, 12(1):1–12.

[4] Jones, I., Wheadon, C., Humphries, S., and Inglis, M. (2016). Fifty years of a level mathematics: Have standards changed? *British Educational Research Journal*.

[5] Lawson, D. (2003). Changes in student entry competencies 1991–2001. *Teaching Mathematics and its Applications*, 22(4):171–175.

[6] Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Kontakt: Dr. Matthew Inglis, Mathematics Education Centre, Loughborough University, Loughborough, Leicestershire LE11 3TU, England. m.j.inglis@lboro.ac.uk

V. l. n. r.: Ian Jones, Christopher Wheadon, Sara Humphries und Matthew Inglis

Dr. Ian Jones ist *Lecturer* am *Mathematics Education Centre* der *Loughborough University* und wissenschaftlicher Berater von *No More Marking Ltd.*

Dr. Christopher Wheadon ist der Gründer von *No More Marking Ltd.* und ist Berater in mehreren nationalen Projekten des *Office of Qualifications and Examinations Regulations* (Ofqual).

Sara Humphries M.Sc. hat Mathematik an der *University of Liverpool* und Psychologie an der *Nottingham Trent University* studiert. Sie ist nun Doktorandin am *Mathematics Education Centre* der *Loughborough University*.

Dr. Matthew Inglis ist *Reader* am *Mathematics Education Centre* der *Loughborough University* und *Honorary Research Fellow* am *Learning Sciences Research Institute* der *University of Nottingham*. Im Jahre 2014 erhielt er den *Annie and John Selden Prize* der *Mathematical Association of America* für seine Forschungsbeiträge zur Didaktik der Universitätsmathematik.

Aus dem Englischen übertragen von Benedikt Löwe unter Mitwirkung von Daniel Sommerhoff.