Peter Birkel

Beurteilungsübereinstimmung bei Mathematikarbeiten?

Zusammenfassung:

131 Lehrer(innen) beurteilten vier Mathematikarbeiten von Viertklässlern. Alle Lehrer verwendeten Punktesystem zur Auswertung. Der relative Anteil der erteilten Punkte konnte bei derselben Arbeit um bis zu 43 Prozentpunkte schwanken. Die abgegebenen Beurteilungen für dieselbe Arbeit differierten um bis zu 2.75 Notenstufen. Damit erscheint die Beurteilung von Mathematikarbeiten als nicht viel reliabler als die Aufsatzbeurteilung. Eine der verwendeten Arbeiten war im Original besonders unsauber in der Ausführung. Von dieser wurde eine "geschönte" Version erstellt und etwa der Hälfte der Beurteiler anstelle der Originalarbeit zur Beurteilung vorgelegt. Bei dieser Arbeit zeigte sich, dass die Auswertung der Arbeit von der Akkuratesse ihrer Ausführung nicht abhängig war, wohl aber die Beurteilung. Die von den Lehrern verwendeten Punktesysteme bei der Auswertung streuten zwischen 16.5 und 55 Punkten. Es zeigte sich, dass die Lehrer mit höchsten maximalen Punktsystemen den Schülern auch mehr Punkte gaben und entsprechend auch bessere Noten.

Summary:

131 teachers scored the exercises of four pupils in mathematics which were originally written by forth graders of primary school. All teachers used scoring systems of different points. The relative amount of points given to the same exercise differed until 43 points on the 100 %-scale. Marks given for the same exercise differed until 2.75 intervals on the 6-point-scale of marks used in Germany. So grading math exercises was not much more reliable than grading essays. One of the four exercises was rather slipshod in the original version. About half of the teachers had to score the original and the others a neat copy written by a student together with the other exercises. It could be shown that the neatness did not affect the measuring of the achievement, but influenced the teacher's marks. Measuring the achievement teachers used scoring systems, which differed from 16.5 to 55 points. It could be shown that teachers using systems with highest maximum of points gave a higher percentage of points and better grades to the pupils.

1 Anlass für die Durchführung der Replikaktionsstudie

Wenn man bei Vorträgen Lehrerinnen und Lehrern der verschiedenen Schularten etwas über die Problematik der Notengebung berichtet und sich dabei auf die Literatur zu diesem Themenbereich stützt, dann sieht man sich u.U. schnell dem Vorwurf ausgesetzt, dass die Beispiele, die man für die mangelnde Reliabilität und Validität der Ziffernbenotung anführt, inzwischen ziemlich alt seien. Ingenkamp hat sein Buch über „die Fragwürdigkeit der Zensurengebung“ in der ersten Auflage bereits 1971 herausgegeben, 1976 ergänzt und 1995 unverändert in Neuauflage erscheinen lassen. Die darin publizierten Originalarbeiten sind noch viel älter und stammen z.T. aus den angloame-

rikanischen Ländern. Von Studierenden und Lehrern werden immer wieder Zweifel gehegt, ob die dort berichteten Ergebnisse noch Relevanz für die heutige Schulsituation in Deutschland haben. Wurde in den vergangenen dreißig bis vierzig Jahren die Validität solcher Untersuchungen nicht einfach dadurch reduziert, dass immer mehr Lehrer bereits während ihrer eigenen Ausbildung ausführlich über diese Problematik informiert wurden? Haben denn nicht in der jüngeren Vergangenheit immer mehr Lehrkräfte Möglichkeiten zur Objektivierung der Leistungsmessung kennen gelernt (s. z.B. Gaude & Teschner 1985, Wendeler 1974)?
Wie sieht es aber um die Ausbildung der Lehrer im Bereich Leistungsmessung und Leistungsbeurteilung aus? Was lernen Lehramtsanwärter heute wirklich an den Hochschulen, um ihre Notengebung auf eine zuverlässigere Basis zu stellen? Aus den Erfahrungen an einer Pädagogischen Hochschule, die Lehrer für die Grund-, Haupt- und Realschule ausbildet, kann berichtet werden, dass von Zeit zu Zeit immer wieder Lehrveranstaltungen zum Problembereich Notengebung angeboten werden, wenngleich auch nicht unbedingt vom Fach Mathematik. Auch in den gängigen Standardwerken zur Mathematikdidaktik findet man nur in seltenen Fällen Hinweise darauf, wie man als Lehrer das Problem Notengebung sinnvoll lösen kann. Eine lobenswerte Ausnahme bildet da vielleicht die Mathematikdidaktik von Leuders (2003). Kriterien zur Einschätzung mathematischer Schülerleistungen oder Teilleistungen wurden bisher nicht entwickelt oder nicht hinreichend publik gemacht.

Aufgrund der Wahlfreiheit bei der Zusammenstellung von Stundenplänen durch die Studierenden werden solche Lehrveranstaltungen zur Problematik der Zensurengebung nur von einem Bruchteil der Lehramtsstudierenden tatsächlich besucht. Neben vielen objektiv vorhandenen Schwierigkeiten bei der Stundenplanzusammenstellung kommt vor allen Dingen ein wesentliches Problem hinzu: Wer sich über die Problematik der Ziffernbenotung informieren will, muss in der Lage sein, empirische Untersuchungen mit all ihren statistischen Angaben zu lesen und zu verstehen. Eine Einführung in die Statistik belegen Lehramtsstudenten aber nur in seltenen Ausnahmefällen. Hinzu kommt, dass die Lehramtsstudierenden ebenso wie die bereits tätigen Lehrkräfte glauben, dass man die Notengebung nicht besonders erlernen muss, weil die doch ganz einfach so fortgeführt werden müsse, wie man sie schon früher als Schüler bei seinen eigenen Lehrerinnen und Lehrern erlebt hat. Diese Erfahrungen stützen die Hypothese, dass so alte Forschungsergebnisse, wie z.B. die zur Problematik der Beurteilung von Mathematikarbeiten, vermutlich auch heute noch ihre Gültigkeit haben.

Solche Überlegungen waren dann der Anlass, die Beurteilung von Mathematikarbeiten ähnlich wie bei Weiss (1966, 1995a) erneut zu untersuchen, zumal seither keine Replikaktionsstudien publiziert wurden. Würden sich ähnliche Ergebnisse wie damals wiederfinden lassen? Im Rahmen einer Lehrveranstaltung an der PH Weingarten wurde das von den Studierenden heftig angezweifelt. Bei der Bewertung eines Aufsatzes konnten sich die Studierenden durchaus vorstellen, dass die Beurteilungen der Lehrer stark divergieren, aber bei einer Mathematikarbeit, bei der doch objektiv feststellbar sein müsse, ob ein Ergebnis richtig oder falsch sei, müssten doch die Lehrer zu übereinstimmenden Ergebnissen kommen! Daraus erwuchs dann der Wunsch, eine solche Untersuchung in der heutigen Zeit einmal zu wiederholen. An die Replizierbarkeit der Ergebnisse von Weiss (1966) glaubten die Studenten von vornherein nicht.

Da der Autor als Veranstaltungsleiter im Fach Pädagogische Psychologie arbeitet, konnte es auch nicht Sinn und Ziel einer solchen Untersuchung sein, Handlungsalternativen für die Lehrkräfte in der Schule zu entwickeln, weil dazu ein erhebliches Ausmaß an fachdidaktischer Kompetenz nötig wäre. Es wäre schon ein Erfolg einer solchen Arbeit, wenn sich Mathematikdidaktiker dadurch angesprochen fühlten, solche Handlungsalternativen zu entwickeln, vielleicht auch in Zusammenarbeit mit entsprechenden Fachleuten aus dem Bereich der Erziehungswissenschaft. Allerdings wären solche Alternativen sicher nicht nur durch vorschnelle Vorschläge zu finden, sondern nur durch einen langwierigen Prozess intensiver, sich wechselseitig befruchtender hermeneutischer und empirischer Forschung. Erstaunlicherweise waren die Ansätze z.B. von Gagné (1965, 1969) zum hierarchischen Aufbau von Wissenssystemen entweder nicht zielführend oder sie wurden nicht rezipiert.

2 Stand der Forschung

Die älteste Untersuchung zur Frage der Übereinstimmung der Lehrer bei der Beurteilung von Mathematikarbeiten datiert bereits aus dem Jahr 1913. Es handelt sich um die bei Ingenkamp (1995) in Übersetzung abgedruckte Untersuchung von Starch & Elliot zur "Verlässlichkeit der Zensuren von Mathematikarbeiten". Damals stellte sich heraus, dass die Lehrer eine im Original kopierte Mathematikarbeit zu einem geometrischen Problem extrem unterschiedlich bewerteten. Umgerechnet auf die in Amerika übliche 100-Punkte-Skala reichten die Bewertungen von 23 bis 92 Punkten. Die am häufigsten gewählte Bewertung waren 75 Punkte, denn das war gerade die Grenzmarke zwischen "bestanden" und "nicht bestanden" und damit der "weder-noch-Punkt", bei dem der Lehrer einer wirklichen Entscheidung aus dem Wege gehen konnte, wenn er nicht recht wusste, wie er die Leistung richtig einordnen konnte.

In den 60er-Jahren beschäftigte sich Weiss im deutschsprachigen Raum mit dem Problem der Urteilsübereinstimmung bei Rechenarbeiten. Er ließ je eine Rechenarbeit der 4. und 5. Klasse von einer größeren Anzahl von Lehrern beurteilen und stellte fest, dass sich die Zensuren praktisch über die ganze Notenskala verteilten. Hinzu kam noch, dass er den beurteilenden Lehrern gezielt positive oder negative Vorinformationen über die Schüler mitteilte, die vermeintlich die Arbeiten geschrieben hatten. Bei der Rechenarbeit die der 4. Klasse entstammte, erwiesen sich die Zensuren als deutlich durch die Vorinformation beeinflusst, während bei der Arbeit des Fünftklässlers dieser Effekt nicht signifikant war.

1971 stellte Haecker die subjektiven Faktoren im Leistungsurteil der Lehrer dar. In einer Befragung von Lehrkräften fand er heraus, dass bei der Beurteilung von Mathematikleistungen recht unterschiedliche Gesichtspunkte eine Rolle spielten. Einige Lehrkräfte nehmen die Zuordnung zur Zensur vor aufgrund der gefundenen Fehler, andere aufgrund der richtigen Lösungen. Die meisten Lehrkräfte schätzten die Schwierigkeit von Aufgaben als leicht, wenn sie einfache Rechenfertigkeiten oder Kopfrechnen erforderten, und als schwer ein, wenn das Finden eines Lösungsansatzes bei Textaufgaben oder mathematisches Denken gefordert waren. Bei der Umwandlung der erteilten Punkte in Zensuren arbeiteten etwa 40 % der Lehrer nach einem festen Schema, während andere die Verteilung der Zensuren vom Abschneiden der Klasse insge-

samt abhängig machten. Bei 10 % ging auch die Bewertung der äußeren Form (Sauberkeit, Schrift) in die endgültige Bewertung ein. Etwa 80 % bekundeten die Bedeutung der Zeugnisnoten vorheriger Lehrer für die eigene Notengebung.

Ähnlich wie Haecker, nur empirisch fundierter, weist Schnotz 1971 auf das Problem der Lehrer hin, den Schwierigkeitsgrad von Aufgaben richtig einschätzen zu können. Meist unterschätzten sie die Schwierigkeit. Große Diskrepanzen bei der Schwierigkeitseinschätzung traten bei Aufgaben aller Schwierigkeitsgrade auf. Wenn bei einzelnen Aufgaben einmal höhere Übereinstimmung bei den Lehrkräfte herrschte, dann nur aufgrund der gemeinsamen Unterschätzung des wirklichen Schwierigkeitsgrades.

1973 legte Dicker eine Diplomarbeit vor, in der er sich eigentlich für Referenzeffekte bei der Beurteilung von unterschiedlichen Zusammenstellungen von Mathematikaufgaben in einer Klassenarbeit interessierte. Gleichzeitig ließ er aber auch Lehrer nach einem Vierteljahr Teile der Arbeiten ein zweites Mal beurteilen. Die Beurteilungsdifferenzen erwiesen sich als auf dem 5 %-Niveau signifikant. Der von Dicker berechnete Reliabilitätskoeffizient für die wiederholte Benotung lag mit r = .46 (N^1 = 24) nur knapp über der Signifikanzgrenze, wies aber doch auf eine deutlichere Übereinstimmung der Urteile hin, als das bei aufsatzähnlichen Aufgaben (Eells 1930/1995, r = .25) der Fall war. Danach wurden keine Forschungsergebnisse zur Reliabilität der Zensurengebung im Fach Mathematik mehr berichtet.

Die fachdidaktischen Erörterungen in der jüngeren Zeit beschäftigen sich eher mit pragmatischen Ansätzen zu Absprachen und Transparenz der Beurteilungskriterien unter Zusammenarbeit des Kollegiums (Kirk 1997), Erfahrungen mit gegenseitiger Beurteilung durch Schüler (Engel 1995), der Funktion von Probearbeiten in der Grundschule (Bauer 1993), allgemeinen Anleitungen zur Leistungsbeurteilung in Zeugnissen (Bartnitzky 1989), zum Zusammenhang von Leistungsbeurteilung und Differenzierung des Mathematikunterrichts (Krampe 1980, Mittelmann & Krampe 1987) und gipfeln in dem augenzwinkernden Vorschlag, die Zensuren vielleicht doch zu würfeln (Herget 1996).

3 Untersuchungsmaterial

Im Rahmen des Tagespraktikums der Lehramtsstudenten an der PH Weingarten suchten die Studierenden aus dem Klassensatz einer Mathematikarbeit, die zufällig kurz vor Beginn des Praktikums im WS 00/01 von Schülern im ersten Halbjahr der vierten Grundschulklasse geschrieben worden war, vier Arbeiten heraus, die man im Rahmen der Untersuchung verwenden wollte. Aufgabe der Schüler war es gewesen, zunächst die Grundrechenarten Addition, Subtraktion und Multiplikation auszuführen und dann einige Umwandlungsaufgaben aus dem Bereich Längenmaße zu bewältigen (s. Anhang A), bevor mit den Aufgaben 8 bis 12 Sachaufgaben zu lösen waren, bei denen die Kinder z.T. auch die Fragen selbst finden mussten (Aufgaben 9 bis 11). Die Klassenlehrerin sah diese Arbeit insgesamt als recht erfolgreich in ihrer Klasse an.

---

[1] Mit N bezeichnet man immer die Anzahl der Fälle, hier also die Anzahl der Lehrer, deren Urteil in die Berechnung einging.

Die für die Untersuchung auszuwählenden Arbeiten sollten sich hinsichtlich der Qualität deutlich unterscheiden, was an den ursprünglich gegebenen Noten abgelesen wurde. Es wurde nicht darauf geachtet, ob die Arbeiten sonderlich „ordentlich" in der Ausführung waren. Besonders bei der Arbeit A fielen viele Durchstreichungen, Überschreibungen und Striche auf dem Rand auf[2]. Eine Studentin stellte von dieser Arbeit eine schön geschriebene und saubere Version her. Damit sollte zusätzlich untersucht werden, ob die äußere Form dieser Arbeit die Beurteilung der gezeigten Leistung beeinflusst.

4 Untersuchungsansatz

Jeder Lehrer, der an der Untersuchung teilnahm, sollte alle vier Mathematikarbeiten beurteilen. Um nun eine einigermaßen systematische Variation der Variable „äußere Form der Arbeit A" zu erreichen, wurde diese Arbeiten in 56 Fällen in schön abgeschriebener Form und in 75 Fällen in der "schlampigen" Originalform beurteilt. Unterschiede in der Fallzahl gehen zurück auf das unterschiedliche Geschick der Studierenden, Lehrkräfte zur Korrektur der Arbeiten zu motivieren.

Gruppe	1	2	
Schüler A	schön[3]	orig.	
Schüler B	orig.	orig.	
Schüler C	orig.	orig.	
Schüler D	orig.	orig.	
N =	56	75	**Σ = 131**

Tab. 1: Design der Untersuchung

5 Stichprobe

Im Rahmen einer Lehrveranstaltung zu Problemen der Schülerbeurteilung wurden Studierendengruppen damit beauftragt, heimatortnah einige Grundschullehrkräfte zu bitten, die vorliegenden vier Arbeiten möglichst mit ausführlicher Begründung der erteilten Noten zu beurteilen.[4]

Die eigentliche Untersuchung fand im Frühjahr 2001 im weiteren Umfeld der PH Weingarten zwischen Bodensee und Donau statt. Insgesamt waren schließlich die Beurteilungen von 131 Lehrerinnen und Lehrern zusammen gekommen. Sie verteilten sich auf die beiden Beurteilungskonstellationen wie aus Tab. 1 zu ersehen ist. Etwa in 42.7 % der Fälle beurteilten die Lehrkräfte die Arbeiten mit der "geschönten" Version der Arbeit A und in 57.3 % der Fälle die Kombination der Originalarbeiten.

---

2 Das Original dieser Arbeit befindet sich im Anhang B.

3 schön = von Studentin schön abgeschrieben, orig. = Kopie der Originalarbeit des Schülers

4 Den beteiligten Studenten winkte als Lohn für die Mühe der Erwerb eines Leistungsnachweises für die Zwischenprüfung.

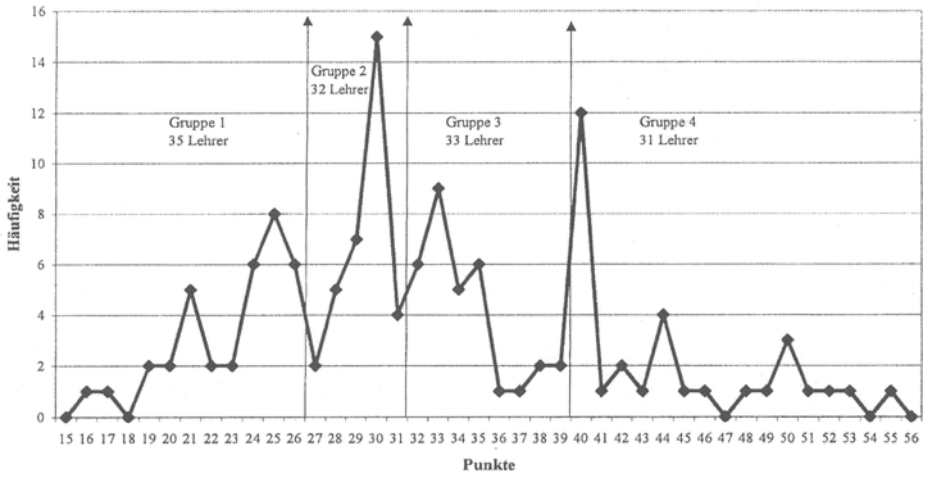
Auf eine Aufgliederung nach Geschlecht und Alter der beteiligten Lehrkräfte musste verzichtet werden, da absolute Anonymität zugesichert war. Diese Informationen wurden nicht erfragt.

6 Beurteilungsverfahren der Lehrer

Eine Überprüfung des Beurteilungsverfahrens der Lehrkräfte erbrachte das übereinstimmende Ergebnis, dass alle die Schülerleistungen in der Mathematikarbeit zunächst nach einem Punktesystem bewerteten, bei dem vermieden wurde, sich nur an der Richtigkeit oder Falschheit des Ergebnisses zu orientieren. Vielmehr analysierte man, wie viele und welche Rechenschritte zur Lösung der gestellten Aufgaben nötig waren. Manchmal gewichteten die Lehrkräfte einzelne Rechenschritte auch noch nach dem geschätzten Schwierigkeitsgrad mit mehreren Punkten. Das war vor allem bei den Sachaufgaben der Fall.

Dadurch bedingt konnte die Maximalzahl der zu vergebenden Punkte sehr unterschiedlich sein. So reichte die Spannweite der zu vergebenden Punkte für die 12 Aufgaben von 16.5 bis zu 55 Punkten. Relativ gesehen wurden 30- bzw. 40-Punkte-Systeme (vgl. Abb. 1) am häufigsten verwendet[5]. Im Prozess der Notengebung war damit der Aspekt der Leistungs"messung" erledigt.

Abb. 1: Vergebene Punkte *(bei halben Punkten abgerundet)*



Der zweite Teilprozess beinhaltet die eigentliche Leistungs"beurteilung", das heißt die Übersetzung der erreichten Punkte in die Notenskala. Hier geht es um die Würdigung der gezeigten Leistung, die die unterschiedlichsten Funktionen (s. dazu die Ausführungen von Dohse, 1995 und Weiss, 1995b) in der Regel gleichzeitig erfüllen soll.

---

5 Die in Abb.1 vorgenommene Gruppeneinteilung dient dazu, a posteriori den Zusammenhang zwischen der Wahl der maximal zu vergebenden Punktzahl und der Notenvergabe zu untersuchen.

So objektiv der Lehrer auch die Leistungsmessung zu gestalten versucht, bei der Leistungsbeurteilung kommt auf jeden Fall wieder eine subjektive Komponente ins Spiel, denn es liegt z.T. im Ermessen der Lehrkraft, welche Modelle zur Übersetzung der Punkteverteilung in Noten sie bevorzugt, ob sie überhaupt ein verbindliches Modell wählt, oder welche zusätzlichen Aspekte (z.B. Lob, Motivation) sie in die Leistungsbeurteilung einfließen lassen will. Viele Lehrer versuchten der Subjektivität dieses Teilprozesses dadurch auszuweichen, dass sie anhand gängiger Tabellen eine lineare Übersetzung der Punkte in Noten vornahmen, bei der ein Intervall gleichmäßig so an die Punkteverteilung angepasst wurde, dass es überall auch die gleiche Punktezahl abdeckte. So deckte z.B. das Notenintervall einer Notestufe auf der 50-Punkte-Skala jeweils immer gerade 10 Punkte ab. Einige Lehrer verwendeten das Programm "WinNote" am PC, andere benutzten eine Formel zur linearen Transformation der Punkte in Noten:

$$Note = 6 - 5 * \frac{erreichtePunktzahl}{\max.Punktzahl}$$

7 Ergebnisse

7.1 Beurteilung der vier Arbeiten

7.1.1 Prozentanteil gegebener Punkte (Aspekt Leistungsmessung)

Aufgrund der Tatsache, dass die Lehrer bei der Auswertung der Mathematikarbeiten sehr unterschiedlich differenzierte Punktesysteme verwendeten, war zu überlegen, wie man die jeweils erteilten Punkte für die Arbeiten vergleichbar machen konnte. Wir entschieden uns dafür, als Vergleichswert die erteilte Punktzahl zur maximalen Punktzahl in Beziehung zu setzen und sie als Prozentsatz zu interpretieren. Die so berechneten Vergleichswerte wurden als abhängige Variable in einer zweifaktoriellen Varianzanalyse mit den Faktoren A (die vier Arbeiten), B (maximale Punktzahl der Punkteskala[6]) verrechnet.

Der resultierende *F-Wert* für den Faktor A (s. Tab. 2) ist mit *720.19* sehr groß und signalisiert eine hohe Signifikanz. Dieser Faktor allein klärt bereits *78.44 %* der Gesamtvarianz. Man kann also mit Fug und Recht sagen, dass die vier Arbeiten sehr unterschiedlich ausgefallen waren.

Aus Abbildung 2 ist zu ersehen, dass die Schüler A bis D prozentual unterschiedlich viele Punkte erreichten. Der beste Schüler A bekam im Schnitt 84 % der jeweils maximalen Punktzahl zuerkannt, während es bei Schüler D im Schnitt nur noch knapp 54 % waren. Die Überprüfung der Unterschiede im Abschneiden mit dem Newman-Keuls-Test (Prüfstatistik q)[7] zeigte, dass die Unterschiede von Schüler zu Schüler jeweils hoch signifikant sind.

---

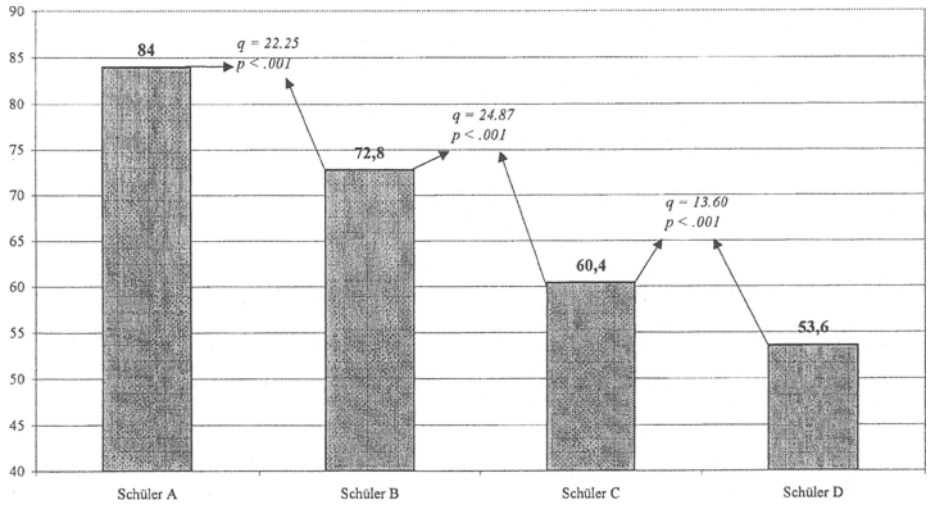
[6] Die Gruppeneinteilung für diesen Faktor wird in Kap. 7.3.1 erläutert.

[7] zur Prüfstatistik q des Newman-Keuls-Tests siehe Heller & Rosemann 1974, S. 205 ff.

Quelle	df[8]	Varianz	F[9]	p <[10]	Effekt %[11]
A (Arbeit)	3	238138.13	720.19	.001	78.44
B (Punkteskala)	3	6366.93	19.26	.001	2.10
A * B	9	1089.96	3.30	.001	1.08
Fehler	507	330.66			18.38
Total	522				100.00

Tab. 2: Ergebnis der Varianzanalyse "Anteil gegebener Punkte"[12]

Abb. 2: Prozentanteil gegebner Punkte



7.1.2 Zensurenmittel für die vier Arbeiten (Aspekt Leistungsbewertung)

Die nächste Varianzanalyse (s. Tab. 3) wurde berechnet mit den von den Lehrern abgegebenen Beurteilungen in Form von Zensuren. Zwischenzensuren wurden nach der in Baden-Württemberg üblichen Verwendung z. B. als 2.75 (für die 3+), 3.25 (für die 3-) oder 3.5 (für die 3-4) verrechnet. Diese Codierungen entsprachen auch den linearen Zensierungsmodellen, die viele Lehrkräfte verwandten.

---

8 *df* = degrees of freedom (Freiheitsgrade), s. dazu z.B. Heller & Rosemann 1974, S. 126.

9 Die Prüfstatistik *F* errechnet sich als Quotient aus der jeweiligen Varianz und der Fehlervarianz.

10 Das Signifikanzniveau wird als Irrtumswahrscheinlichkeit p < angeben (s. Heller & Rosemann 1974, S. 187f).

11 Hier handelt es sich um ein Relevanzmaß. Ein Effekt kann statistisch sehr wohl hochsignifikant sein (z.B. der der Interaktion A*B), aber wenig praktische Relevanz besitzen, da nur gerade 1% der Varianz klärt.

12 Alle statistischen Analysen wurden mit dem Programmpaket KMSS-7 von Kleiter (2002) berechnet.

Nach der Varianzanalyse für die prozentuierten Punktwerte ist zu erwarten, dass auch hier die vier Arbeiten sich bezüglich der Zensuren[13] signifikant unterscheiden.

Erwartungsgemäß unterscheiden sich die durchschnittlichen Beurteilungen der vier Arbeiten ebenfalls hoch signifikant auf der Zensurenskala. Der Faktor A klärt erneut mit *69.24 %* einen ganz erheblichen Anteil an der Gesamtvarianz, wenngleich der auch um fast *9 %* niedriger liegt als bei den Auswertungsergebnissen. Das spricht dafür, dass bei der Beurteilung der Leistung zusätzliche (leistungsfremde?) Faktoren eingeflossen sein könnten.

Quelle	*df*	*Varianz*	*F*	*p <*	*Effekt %*
A (Arbeit)	3	6515.85	427.61	.001	69.24
B (Punkteskala)	3	178.05	11.69	.001	1.89
A * B	9	22.52	1.48	n.s.	0.72
Fehler	507	15.67			28.15
Total	522				100.00

Tab. 3: Ergebnis der Varianzanalyse "Zensuren"

Trotzdem kann man sagen, dass die unterschiedliche Qualität der Leistungen, die sich in den vier Arbeiten dokumentiert, im Wesentlichen für die Unterschiede in den Zensuren verantwortlich ist. Das bedeutet, dass zumindest die Rangfolge bei der Beurteilung der Arbeiten zum größten Teil eingehalten wurde. Inwieweit allerdings Übereinstimmung bei den gegebenen Zensuren herrscht, wird noch weiter untersucht werden.

Betrachtet man die in Abb. 3 auf der nächsten Seite dargestellten Zensurenmittel, so stellt man fest, dass die Arbeit A in Schnitt etwa mit einer 2+ beurteilt wurde. Arbeit B erhielt in etwa eine 2-3, Arbeit C ziemlich genau eine 3 und Arbeit D eine 3-4. Wie bei den prozentuierten Punkten sind auch bei den Zensurenmitteln die Notendifferenzen der "benachbarten" Arbeiten bereits hoch signifikant unterschiedlich.

Am Rande sei erwähnt, dass die Durchschnittsbeurteilungen recht genau den ursprünglich erteilten Noten entsprechen.

7.2 Übereinstimmung der Lehrerurteile

7.2.1 Auswertung im Punktesystem der Lehrer

Als nächstes war von Interesse, inwieweit die Lehrer bei der Auswertung der Arbeiten übereinstimmten. Erwartung zumindest der Studierenden in Weingarten war, dass doch

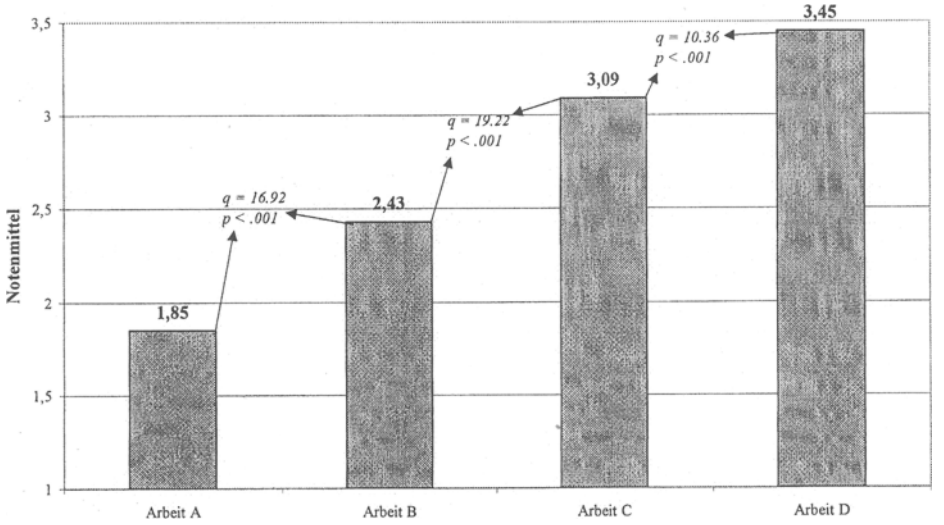
---

[13] Gegen eine Anwendung der Varianzanalyse bei Zensuren könnte die fragliche Skalenqualität als Argument vorgebracht werden. Dieses Argument wird aber einerseits durch den Hinweis relativiert, dass es sich bei der Varianzanalyse um ein robustes Verfahren handelt, das auch gewisse Verstöße gegen die Voraussetzungen bezüglich der Skalenqualität toleriert (s. dazu z.B. Keppel & Saufley 1980, S. 97), und andererseits durch Überprüfungen von Notenverteilungen mittels Standardnormalverteilung (s. z.B. Birkel 1978, S. 240), nach denen immerhin näherungsweise Äquidistanz im mittleren Bereich der Notenskala unterstellt werden kann.

alle Lehrer mehr oder weniger zum gleichen Ergebnis kommen müssten, weil doch bei einer Mathematikarbeit "objektiv feststellbar" sei, wie viel ein Schüler jeweils richtig bearbeitet habe.

Der Blick auf Abb. 4 (nächste Seite) zeigt allerdings, dass bezüglich der prozentuierten Punktwerte große Unterschiede auftraten. So reicht der Anteil der erreichten Punkte bei Schüler A von etwa 69 % bis 95 %! Mit anderen Worten, ein Lehrer war der Meinung, dass dieser Schüler gut ⅔ aller Punkte erreicht hätte, während ein anderer glaubte, dass bei diesem Schüler mit 95 % der möglichen Punkte nur vergleichsweise wenig Punkte abzuziehen waren. Bei der Auswertung der Arbeit mit Hilfe des jeweiligen Punktesystems traten große Differenzen auf, die sich über 26 Punkte auf der Prozentpunkteskala erstreckten.

Abb. 3: Notenmittel der vier Arbeiten



Bei Schüler B (Abb. 5, nächste Seite) reichen die Punktwerte von 52 bis 83 %. Es ergab sich somit eine Differenz von 31 Prozentpunkten! Selbst wenn man hier den "Ausreißer" mit 52 % wegließe, würden die gegebenen Punkte noch von 62 bis 83 % reichen und damit nicht gerade von großer Einigkeit bei der Auswertung der Arbeit zeugen. Bei Schüler C reichen die Werte von 47 bis 73 % und streuen über einen Bereich von wiederum 25 Prozentpunkten.

Bei Schüler D (Abb. 4), dessen Arbeit am schlechtesten bewertet wurde, reichen die gegebenen Punkte von 30 bis 73 %. Sie streuen also über einen Bereich von 43 Prozentpunkten, hier waren die Auswertungsdiskrepanzen der Lehrer am größten! Eine Lehrkraft gab diesem Schüler knapp ein Drittel der zu vergebenden Punkte, während eine andere der Auffassung war, dass derselbe Schüler fast drei Viertel aller möglichen Punkte verdient habe, also mehr als doppelt so viel.

Abb. 4: Gegebene Prozentpunkte bei der besten und schlechtesten Arbeit (Schüler A und D)

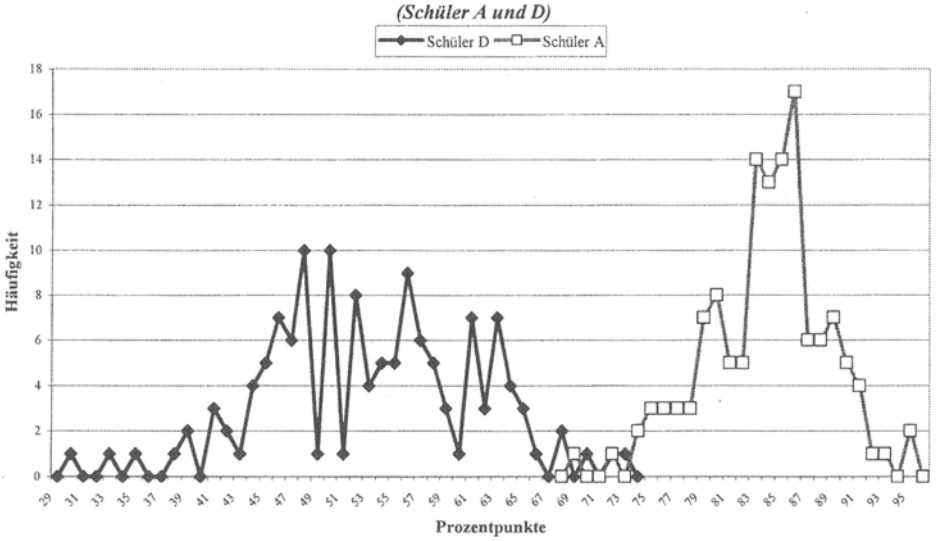
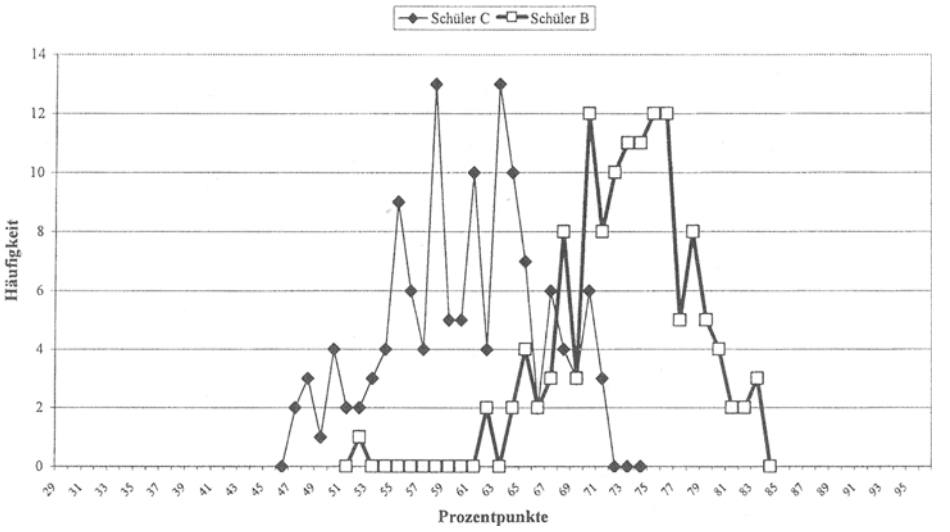


Abb. 5: Gegebene Prozentpunkte für Schüler B und C



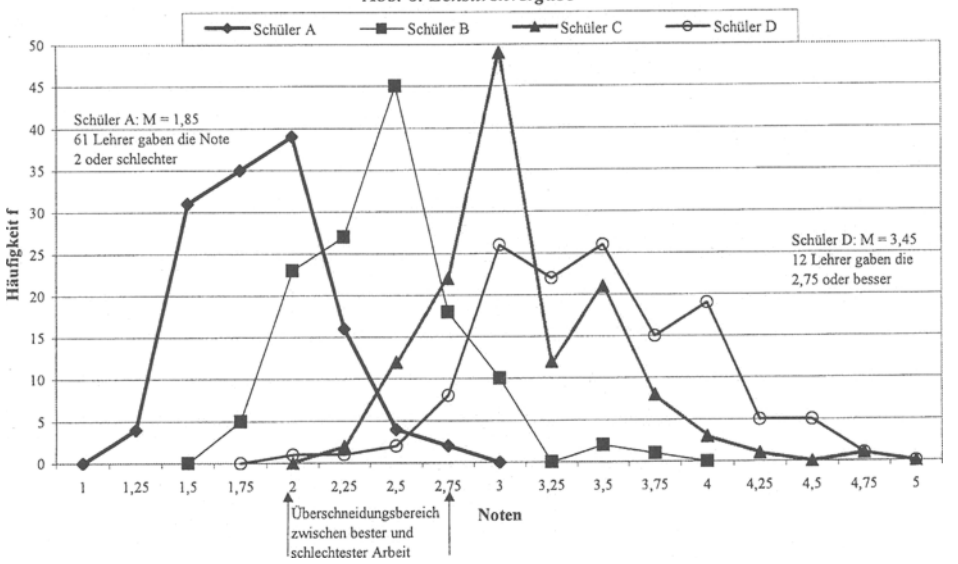
Sind schon die bisherigen Auswertungsunterschiede beträchtlich, so ist man von den Unterschieden bei Schüler D geradezu verblüfft und überlegt sich, wie solche Unterschiede zustande kommen können. Die Studierenden, die an der Sammlung der Daten beteiligt waren, glaubten ja, dass es doch relativ eindeutig entscheidbar sei, inwieweit eine Aufgabe richtig gelöst ist. Diese Ergebnisse belegen, dass die Lehrkräfte doch z.T. ganz unterschiedlich auswerten, ganz unterschiedliche Teilleistungen bewerteten und aufgrund von Schwierigkeitseinschätzungen unterschiedlich viele Punkte vergaben!

Besonders schwer dürfte es fallen zu erklären, dass sich die Bereiche der vergebenen Punktanteile zwischen der besten und der schwächsten Arbeit überschneiden. Es gab Lehrkräfte, die bei der besten Arbeit einmal 69 und einmal 72 % der maximalen Punktzahl gaben, während andererseits bei der schwächsten Arbeit drei Lehrkräfte zu dem Ergebnis kamen, dass 70, 72 oder sogar 73 % der Punkte zu geben seien. Eine Kontrolle der entsprechenden Arbeiten ergab, dass es sich bei diesen drei Lehrkräften um solche handelte, die insgesamt auch die höchsten maximal erreichbaren Punktzahlen zur Auswertung benutzten.

7.2.2 Beurteilung der Leistungen auf der Notenskala

Bei der Beurteilung einer Leistung spielen die vielen möglichen Funktionen eine Rolle, die ein Lehrer bei der Festlegung der Note berücksichtigen will. Darum ist bei den gegebenen Noten durchaus eine gewisse Bandbreite erwartet worden.

Abb. 6: Zensurenvergabe



Aus Abb. 6 ist zu ersehen, dass bei den Beurteilungen auf der Notenskala jeweils deutliche Differenzen auftreten. Bei Schüler A, der die am besten bewertete Arbeit schrieb, reichen die Zensuren von der 1.25 bis zur 2.75. Sie streuen also über 1.5 Notenstufen. Bei Schüler B reichen die Zensuren von der 1.75 bis zur 3.75, eine Streuung über zwei volle Notenstufen. Schüler C erhält im besten Fall eine 2.25 und im schlechtesten Fall eine 4.75. Hier streuen also die Noten über 2.5 Notenstufen. Die größten Urteilsdifferenzen treten bei Schüler D auf. Hier reichen die Zensuren von der glatten 2.0 bis zur 4.75 und umfassen damit sogar 2.75 Notenstufen. Erstaunlicherweise war hier ein Lehrer der Meinung, dass diese Arbeit sogar besser zu bewerten sei als die des Schülers C, der mehr Aufgaben richtig bearbeitet hatte! Hier könnte die unterschiedliche Gewichtung einzelner Aufgaben eine Rolle gespielt haben. Die Spannweite der Beurteilungen kann wohl auch als Ausdruck der unterschiedlichen Funktionen der

Zensuren verstanden werden, die die Lehrkräfte jeweils ganz individuell betonen wollten. Vor allem scheint hier die motivationale Funktion der Notengebung eine Rolle zu spielen, die vor allem die Schwächsten ermutigen und anspornen soll.

Generell lässt sich feststellen, dass die Übereinstimmung bei der Beurteilung der Mathematikarbeiten nicht wesentlich größer ist als die bei der Beurteilung von Aufsätzen (s. Birkel & Birkel 2002, Birkel 2003). Ähnlich wie bei der Untersuchung zur Aufsatzbeurteilung bewirkte auch hier die im Untersuchungsmaterial bereits enthaltene Normierung der Beurteilung durch die in den Arbeiten tatsächlich dokumentierten Leistungen keine deutliche Verringerung der Streubreiten der Beurteilungen. Oder falls das doch der Fall gewesen sein sollte, wie groß wären dann die Streubreiten ausgefallen ohne den Einfluss dieser Normierung? Die von den Studierenden erwartete "große Objektivität" bei der Beurteilung von Mathematikarbeiten ist wohl doch eher Wunsch als Wirklichkeit!

7.3 Die Sauberkeit der Darstellung als Einflussfaktor

Dass die saubere und ordentliche Darstellung einer Klassenarbeit die Beurteilung beeinflussen kann, ist sicher weithin bekannt. Haecker (1971) berichtet, dass etliche Lehrkräfte auch Sauberkeit und Schrift in die Bewertung von Mathematikarbeiten einfließen lassen. Weiss (1966, 1995) konfrontierte zwar seine beurteilenden Lehrer nicht mit entsprechend gestalteten "Originalarbeiten", sondern teilte ihnen nur mit, dass "die Original-Arbeit durch unsaubere Form und schlampige Schrift" (Weiss 1995, S. 108) auffiel. Im Vergleich zu der Arbeit, die als von "einem mathematisch begabten Schüler mit einer Neigung zu originellen Lösungen" (Weiss 1995, S. 108) stammend ausgegeben war, wurde die "schlampige" Arbeit eines Viertklassschülers als signifikant schlechter eingestuft. Bei der Arbeit eines Fünftklässlers erreichte der Unterschied die Signifikanzgrenze nicht ganz. Als Grund dafür könnte u.E. die Tatsache aufgeführt werden, dass diese Arbeit insgesamt deutlich schlechter ausgefallen war, keiner der Lehrer gab hier noch die Note "1". Hier könnte wiederum gegolten haben, dass es "offenbar ... wohl leichter [ist], bei einer schlechten Leistung ein größeres Maß an Übereinstimmung zu erreichen als bei mittelmäßigen oder guten Leistungen" (Birkel & Birkel 2002, S. 222).

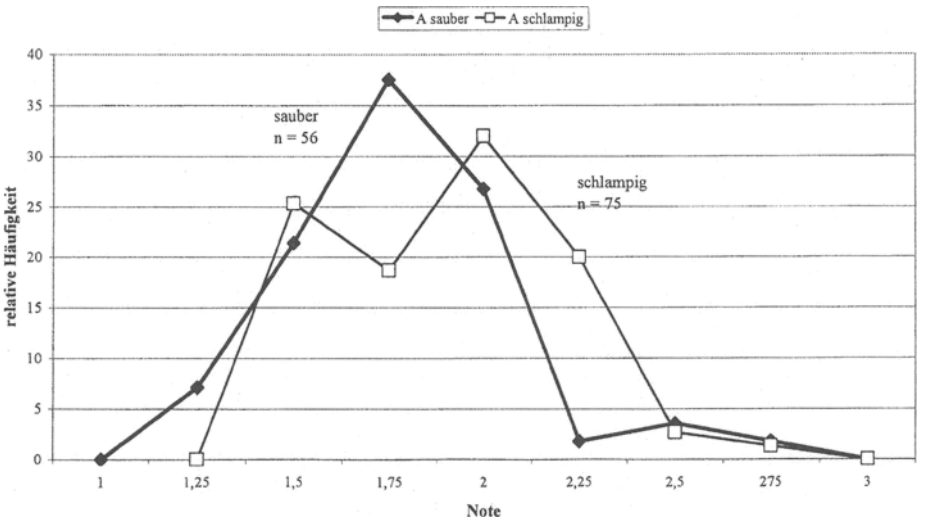
Arbeitete Weiss (1995a) also noch mit Vorinformationen zu den Arbeiten, so wollten wir die Untersuchung ohne verbalen Hinweis auf die Gestaltung der Arbeit durchführen, sondern das Aussehen der Arbeiten direkt wirken lassen. Dazu wurde die "geschönte" Fassung der Arbeit von Schüler A knapp der Hälfte der Lehrer (42.7 %) zur Beurteilung vorgelegt, um überprüfen zu können, inwieweit die äußere Form der Arbeit die Beurteilung der Leistung beeinflusst. Ein Hinweis zur Beachtung der äußeren Form der Arbeit, zur Bildung einer besonderen Note für diesen Aspekt und zur eventuellen Berücksichtigung dieser bei der Gesamtnote wurde nicht gegeben. Bei keiner Lehrkraft tauchte ein besonderer Vermerk zur mangelnden Sauberkeit und Ordnung der Arbeit des Schülers A auf.

Bei dessen Arbeit wurden sowohl die Werte für den Anteil gegebener Punkte als auch die Noten mit einer Einweg-Varianzanalyse geprüft. Es stellte sich heraus, dass der Anteil gegebener Punkte sich als nicht beeinflusst durch die Sauberkeit der Darstel-

lung erwies (*F=0.56, df=1/129, n.s.*). In der schön geschriebenen Version wurden durchschnittlich 84.79 % der Punkte gegeben, in der "schlampigen" Originalversion 83.88 %, also fast genau so viel. Unabhängig von der Tatsache, ob die Lehrer viele oder eher weniger Punkte in ihrem Punkteschema verwendeten, gaben sie der Arbeit des Schülers A praktisch im Schnitt die gleichen prozentualen Punktanteile. Bei den Zensuren als abhängiger Variable wurde der Mittelwertsunterschied allerdings auf dem 5%-Niveau signifikant (*F=4.87, df=1/129, p<.05*). In der schön geschriebenen Version wurde im *Mittel* die Note *1.78* erteilt, in der Originalversion die Note *1.90*.

Man kann also feststellen, dass die Sauberkeit der Darstellung **keinen Einfluss** ausübte **auf die Auswertung** der Arbeit und damit auf den Anteil der gegebenen Punkte. Erst beim Prozess der Notenfindung, also beim **Beurteilungsprozess**, spielte dieser Faktor eine Rolle. In Abb. 7 erkennt man deutlich, dass die Notenverteilung bei sauberer Darstellung deutlich zum positiven Pol hin verschoben ist. Da keinerlei Hinweis auf die Berücksichtigung der äußeren Form bei der Notengebung erfolgt war, und keiner der Lehrer eine entsprechende Bemerkung notiert hatte, dürfte dieser Einfluss eher unbewusst wirksam geworden sein

Abb. 7: Notenverteilung mit sauberer und schlampiger Darstellung



7.4 Zusammenhang zwischen maximaler Punktzahl bei der Auswertung und Beurteilung der Arbeiten

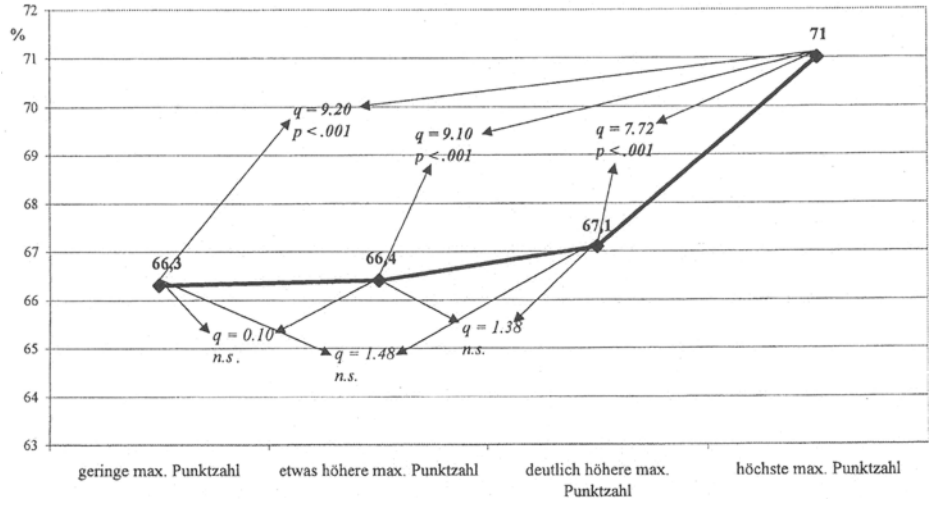
7.4.1 Zusammenhang mit dem Anteil gegebener Punkte

Nach den Erfahrungen in Kap. 6.2.1. wird der Frage nachgegangen, ob nicht vielleicht generell ein Zusammenhang zwischen der Wahl der maximalen Punktzahl der Punkteskala und dem relativen Anteil der gegebenen Punkte bestehen könnte. Geben also u. U. die Lehrer, die eine Punkteskala mit hoher maximaler Punktzahl benutzten, auch

relativ gesehen mehr Punkte als die Lehrer, die nur eine geringere maximale Punktezahl verwendeten? Steckt mit anderen Worten hinter der Entscheidung für die Wahl einer Punkteskala eventuell noch eine andere Motivation als nur die, die Auswertung möglichst transparent zu gewährleisten? Die größere Transparenz des Aspekts Leistungsmessung war nach Auskunft der Lehrer der wichtigste Grund für die Verwendung eines Punktesystems.

Dazu wurden die Lehrer a posteriori in vier etwa gleich große Gruppen eingeteilt (s. auch Abb. 1). Die erste Gruppe bildeten die Lehrer, die die geringsten maximalen Punktzahlen verwendeten. Diese 35 Lehrer vergaben maximal zwischen 16.5 und 26 Punkte. In der nächsten Gruppe verwendeten 32 Lehrer etwas höhere maximale Punktzahlen und vergaben zwischen 27 und 31 Punkte. Als dritte Gruppe wurden die 33 Lehrer zusammengefasst, die mit 31.5 bis 39 Punkten eine noch höhere maximale Punktezahl verwendet hatten. Schließlich bildeten die 31 Lehrer, die maximal 40 oder mehr (bis zu 55) Punkte verwendeten, die vierte Gruppe. Wie trennscharf diese Einteilung sein würde, konnte nicht von vornherein eingeschätzt werden, da nicht geklärt war, inwieweit 20 Punkte mit einem ½-Punkte-Raster mit einer 40-Punkte-Skala mit ausschließlicher Vergabe von ganzen Punkten gleichzusetzen sei. Eine stichprobenartige Überprüfung der Punktesysteme der Lehrer signalisierte allerdings, dass nur in Ausnahmefällen auch mit halben Punkten gearbeitet wurde. Es erhärtete sich der Eindruck, dass mit der Erhöhung der maximalen Punktzahl eher eine differenziertere Gewichtung von Rechenschritten einherging. Besonders deutlich trat das bei der Bewertung der Sachaufgaben hervor, bei denen eine Lehrkraft z.B. nur 2 Punkte vergab, während eine andere bis zu 8 Punkte zur Verfügung stellte.

Abb. 8: Prozent gegebener Punkte



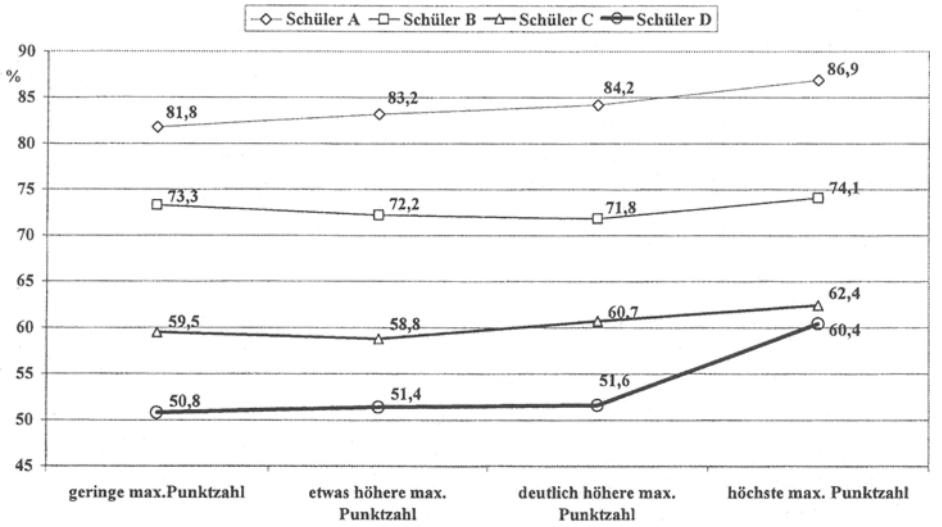
Schaut man sich den Effekt der so eingeteilten Lehrergruppen auf den Anteil der gegebenen Punkte an (s. Tab. 2), so stellt man fest, dass mit *F=19.26 (df=3, p<.001)* dieser Effekt hoch signifikant ist. Aus Abb. 8 ist zu ersehen, dass die Lehrkräfte, die die höchsten maximalen Punktwerte nutzten, den Schülern mit 71 % der möglichen

Punkte einen signifikant höheren Anteil der Punkte gaben als die Lehrer der anderen Gruppen. Eine Überprüfung der Mittelwertsunterschiede mit dem Spannweiten-Test nach Newman-Keuls (s. Heller & Rosemann 1974) ergab, dass die Unterschiede zwischen den ersten drei Gruppen nicht signifikant waren, auch wenn hier bereits ein geringer Anstieg der Werte erkennbar war, diese sich aber jeweils signifikant von der vierten Gruppe mit den höchsten maximalen Punktwerten unterschieden.

Man könnte vermuten, dass Lehrkräfte, die so hohe maximale Punktwerte verwenden, damit eine Absicht verfolgen, nämlich die, möglichst viele Teilschritte im Rechenprozess als gelungen oder vom Denkansatz her als richtig einstufen zu können. Entsprechend viele der möglichen Punkte können dann gegeben werden.

Wie aus Tab. 2 ersichtlich wird neben dem Haupteffekt dieses Faktors auch noch die Interaktion A*B signifikant ($F = 3.30$, $df = 9/507$, $p < .001$). Das deutet darauf hin, dass nicht alle Schüler in gleicher Weise von der Tendenz der Lehrer profitierten, höchste maximale Punktwerte zu verwenden. Um hier die Verhältnisse zu verdeutlichen, seien die durchschnittlichen Anteile gegebener Punkte jeweils für die vier Schülerarbeiten getrennt dargestellt.

Abb. 9: Auswirkung der Höhe der max. Punktzahl bei den einzelnen Schülern



Aus Abb. 9 wird nun ersichtlich, dass tatsächlich nicht alle Schüler in gleicher Weise von der Tendenz profitierten, höchste maximale Punktzahlen zu verwenden. Zwar bekommen die Schüler A bis C auch jeweils ein um etwa 2 Punkte besseres Ergebnis, diese Zugewinne bewegen sich allerdings im Bereich der Zufälligkeit und verfehlen das Signifikanzniveau von 5 %. Bei dem am schlechtesten beurteilten Schüler D allerdings steigt der Anteil gegebener Punkte von 51.6 auf 60.4 %. Dieser Anstieg um fast 9 % ist hoch signifikant.

Dieses Ergebnis bedeutet, dass die weiter oben bereits ausgeführte Interpretation bezüglich der Absicht der Lehrer, die höchste maximale Punktwerte verwenden, modifiziert werden muss. Zwar profitieren tendenziell alle Schüler ein wenig von dieser Auswertungstendenz, aber der schwächste Schüler profitiert in ganz besonderem Maße.

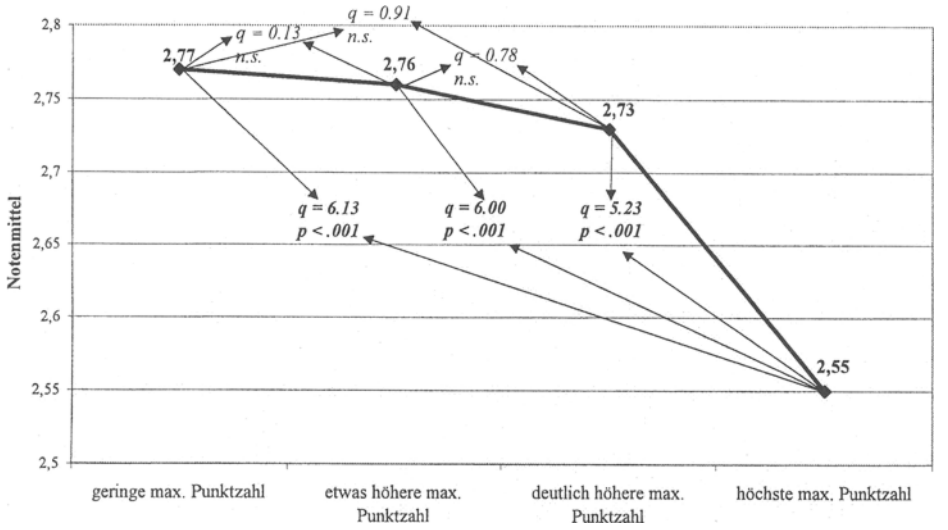
Unsere Vermutungen gehen in die Richtung, dass diese Lehrer vor allem den schwächsten Schüler durch die Rückmeldung motivieren möchten, trotz allem eine vergleichsweise akzeptable Leistung erbracht zu haben. Diese Lehrkräfte schätzen die Leistung von Schüler D als in etwa gleich gut ein wie die von Schüler C, denn der Beurteilungsunterschied zwischen diesen beiden Arbeiten wird hier nicht mehr signifikant.

7.4.2 Zusammenhang mit der Erteilung der Noten

Ähnlich wie der Anteil gegebener Punkte ließ sich natürlich auch die Note als abhängige Variable in einer Varianzanalyse mit dem Faktor "Höhe des max. Punktwertes" verrechnen. Die entsprechenden Ergebnisse sind bei aller Vorsicht wegen der oben bereits erwähnten Einschränkungen bei der Interpretation der Tab. 3 zu entnehmen. Hier geht es jetzt um die Frage, ob die zusätzlich gegebenen Punkte bei Verwendung höchster maximaler Punktwerte auch zu besseren Zensuren führen. Auch dieser Haupteffekt ist mit einem *F-Wert* von *11.69* hoch signifikant ($df = 3/507$, $p < .001$)

Aus Abb. 10 wird unmittelbar deutlich, dass sich die Beziehung zwischen der Tendenz zur Verwendung höchster maximaler Punktzahlen und dem Anteil gegebener Punkte direkt auch in der Beziehung zur Notenskala widerspiegelt. In den drei Gruppen mit weniger hohen maximalen Punkteskalen zeichnet sich eine sehr moderate Verbesserung der Noten ab, die ebenso wie beim Anteil der gegebenen Punkte nicht signifikant ist. Nur die Gruppe der Lehrer, die die höchsten maximalen Punktwerte verwendete, weicht hoch signifikant vom Urteil der drei anderen Gruppen ab und gibt deutlich bessere Noten.

Abb. 10: Notenmittel über alle vier Arbeiten



Die Feststellung dieser Tatsache liefert natürlich noch keine Aufklärung über ihre Ursachen. Es sind vermutlich Überlegungen in vielerlei Richtungen möglich, aber wenn man einmal spekulativ versucht, sich dieses Ergebnis zu erklären, könnte man da

bei aller Vorsicht nicht eventuell so argumentieren? "Die Lehrer verwenden Skalen mit höchsten maximalen Punktzahlen, um den Schülern bessere Noten geben zu können." Die Verwendung von Skalen mit hohen maximalen Punktzahlen könnten als Indiz für eine "schülerfreundliche" Beurteilungstendenz der Lehrer angesehen werden. Positiv gesehen betonen diese Lehrer die motivationale Funktion der Zensur, indem sie die schwächeren Schüler ermuntern, mit ihren Leistungsbemühungen fortzufahren. Dazu hätten sie natürlich den Schüler besser kennen müssen. Ohne konkrete Kenntnis des Schülers könnte dieser die Zensur auch in der Form interpretieren, dass er sich sagt, es sei ja alles in Ordnung mit seiner Leistung, er brauche sich nicht noch zusätzlich anzustrengen.

8 Zusammenfassung

Fasst man noch einmal die wichtigsten Ergebnisse dieser Untersuchung zusammen, so kann man sagen:

1. Die Beurteilung von Mathematikarbeiten ist bei Weitem nicht so zuverlässig, wie es die Studierenden erwartet haben. Zum Teil ergeben sich beträchtliche Urteilsdifferenzen (bis zu 2.75 Notenstufen), die fast die Größenordnung erreichen, wie sie von den Untersuchungen zur Aufsatzbeurteilung her bekannt sind.
2. Trotzdem richten sich die Lehrkräfte bei der Beurteilung der Arbeiten im Wesentlichen nach der unterschiedlichen Leistung der Schüler. Der Faktor A "Arbeit" erreicht immerhin fast 70 % Varianzaufklärung.
3. Die Beurteilung der Arbeit des Schülers A erwies sich als signifikant abhängig von der Akkuratesse ihrer Ausführung. Betrachtet man die Notenfindung als zweistufigen Prozess, so kann man feststellen, dass die Messung der Leistung (Verteilung der Punkte) durch die Darstellungsform nicht beeinflusst war, sondern dass sich die Ausführung der Arbeit nur auf den Beurteilungsprozess (Übersetzung der Punkte in Noten) auswirkte.
4. Die Lehrkräfte verwendeten zur Auswertung der Mathematikarbeiten durchweg Punktesysteme. Es entspricht wohl der Erfahrung der Lehrkräfte, dass ein solches Vorgehen sinnvoll und praktisch ist. Eine solche Maßnahme kann im Prinzip als Beitrag zur "Objektivierung der Leistungsmessung" und zur Erhöhung der Transparenz des Auswertungsprozesses angesehen werden.
5. Die Anzahl der maximal zu vergebenden Punkte schwankte bei den Lehrkräften ganz erheblich. Die verwendeten Punkteskalen reichten von 16,5 bis 55 Punkte. Welche Rechenschritte dabei wie fein unterschieden wurden, hing ganz von den bisherigen Praxiserfahrungen der Lehrkräfte ab, denn konkrete Hinweise dazu wurden und werden im Rahmen der Lehrerausbildung nicht verbindlich vermittelt. Möglicherweise sind manche Punktesystem miteinander vergleichbar, weil ganze oder halbe Punkte verwendet wurden. Hier würde man bei weiteren Untersuchungen genauer nachschauen müssen.
6. Der relative Anteil an gegebenen Punkten konnte sich allerdings beträchtlich unterscheiden. Damit wurde klar, dass die Erwartungen der Studierenden an die Verwendung solcher Punktesysteme enttäuscht wurden, die darauf hinaus

liefen, dass unabhängig von der maximalen Punktzahl der relative Anteil gegebener Punkte vergleichbar sein müsste.

7. Die Verwendung höchster maximaler Punktzahlen ging einher mit einer Erhöhung des Anteils vergebener Punkte und der Vergabe besserer Noten. Davon profitierte der am schlechtesten beurteilte Schüler am meisten. Man könnte das als Indikator für die "Schülerfreundlichkeit" der Beurteilung durch den Lehrer ansehen.

Fazit: Es lässt sich auch heute noch "zeigen, wie absurd die Annahme ist, eine Mathematikarbeit werde mit größerer Präzision beurteilt als eine sprachliche oder irgendeine andere Art von Prüfungsarbeit." (Starch & Elliot 1913, S. 257, zitiert nach Ingenkamp 1975, S. 32)

9 Literatur

Bartnitzky, H. (Hrsg.). [1989]: *Umgang mit Zensuren in allen Fächern. Leistungen und Leistungsförderung, Beobachtungen, Tests, Klassenarbeiten, Zeugnisschreiben.* Frankfurt/M.: Cornelsen Scriptor.

Birkel, P. [1978]: *Mündliche Prüfungen. Zur Objektivität und Validität der Leistungsbeurteilung.* Bochum: Kamp

Birkel, P. [2003]. Aufsatzbeurteilung - ein altes Problem neu untersucht. *Didaktik Deutsch.* 9(2003)15, 46-63.

Birkel, P. & Birkel, C. [2002]: Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? *Psychologie in Erziehung und Unterricht.* 49(2002)3, 219-224.

Bauer, L. [1993]: Hanna - Note sechs. Analysen und Überlegungen zu einer Mathematikprobearbeit der dritten Jahrgangsstufe. *Grundschule*, 25(1993)11, 35-38.

Dicker, H. [1973]: Untersuchung zur Beurteilung von Mathematikaufgaben. Unveröff. Diplomarbeit: EWH Rheinland-Pfalz, Abt. Landau; auszugsweise veröff. in: K. Ingenkamp (Hrsg.). (1995)[9]: *Die Fragwürdigkeit der Zensurengebung.* Weinheim: Beltz, 173-176.

Dohse, W. [1995]: Die Funktionen der Zensur. In: K. Ingenkamp (Hrsg.). *Die Fragwürdigkeit der Zensurengebung.* Weinheim: Beltz, 56-91.

Eells, W.C. [1930]: Reliability of repeated grading of essay type examinations. *Journal of Educational Psychology*, 21(1930)48-52; In Übersetzung abgedruckt in: K. Ingenkamp (Hrsg.). (1995)[9]: *Die Fragwürdigkeit der Zensurengebung.* Weinheim: Beltz, 167-172.

Engel, M. [1995]: Ein Oberstufenkurs benotet sich selbst. Eine Fallstudie. *Schul-Management*, 26(1995)6, 34-37.

Gaude, P. & Teschner, W. [1985]: *Objektivierte Leistungsmessung in der Schule.* Frankfurt/M.: Diesterweg.

Gagné, R. [1965]: The Analysis of Instructional Objectives for the Design of Instruction. In: R. Glaser (Ed.): *Teaching Machines and Programmed Learning II.* Washington DC.

Gagné, R. [1969]: *Die Bedingungen des menschlichen Lernens.* Hannover: Schroedel

Haecker, H. [1971]: Subjektive Faktoren im Leistungsurteil der Lehrer. *Schule und Psychologie*, 18(1971)74-84.

Heller, K. & Rosemann, B. [1974]: *Planung und Auswertung empirischer Untersuchungen.* Stuttgart: Klett.

Herget, W. [1996]: Zensuren würfeln. Wahrlich objektive Zensuren - im Stochastik-Kurs. *Friedrich-Jahresheft, XIV: Prüfen und beurteilen.* (1996)126-127.

Ingenkamp, K. [1975]: *Pädagogische Diagnostik.* Weinheim: Beltz
Ingenkamp, K. (Hrsg.). [1995][9]: *Die Fragwürdigkeit der Zensurengebung.* Weinheim: Beltz.
Keppel, G. & Saufley, W. H. [1980]: *Design and analysis.* San Francisco: W. H. Freeman & Co.
Kirk, S. [1997]: Keine Not mit den Noten? Zur Praxis der Beurteilung von Mathematikarbeiten. *Praxis Schule 5 - 10*, 8(1997)6, 15-17.
Kleiter, E. [2002]: *KMSS-7. Kleiter Mikrocomputer Statistik System - Version 7.* Kiel.
Krampe, J. [1980]: Planung und Auswertung didaktisch-differenzierter Mathematik-Arbeiten. *Sachunterricht und Mathematik in der Primarstufe*, 8(1980)8, 302-307.
Leuders, T. (Hrsg.). [2003]: *Mathematik Didaktik.* Frankfurt/M./Berlin: Cornelsen Scriptor
Mittelmann, R. & Krampe, J. [1987]: Bewertung von Klassenarbeiten. *Mathematische Unterrichtspraxis*, 8(1987)3, 25-28.
Schnotz, W. [1971]: Schätzung von Aufgabenschwierigkeiten durch Lehrer. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 3(1971)2, 106-120
Starch, D. & Elliot, E.C. [1913]: Reliability of grading work in mathematics. *School review*, 21(1913) 254-259 & 280-281. Abdruck in deutscher Übersetzung bei: K. Ingenkamp (Hrsg.). (1995)[9]: *Die Fragwürdigkeit der Zensurengebung.* Weinheim: Beltz, 81-89.
Weiss, R. [1995a]: Die Zuverlässigkeit der Ziffernbenotung bei Aufsätzen und Rechenarbeiten. In: K. Ingenkamp (Hrsg.): *Die Fragwürdigkeit der Zensurengebung.* Weinheim: Beltz, 104-116.
Weiss, R. [1995b]: Aufgaben der Zensuren und Zeugnisse. In: K. Ingenkamp (Hrsg.): *Die Fragwürdigkeit der Zensurengebung.* Weinheim: Beltz, 62-65.
Weiss, R. [1966]: Über die Zuverlässigkeit der Ziffernbenotung bei Rechenarbeiten. *Schule und Psychologie*, 5(1966), 144-151.
Wendeler, J. [1974]: *Standardarbeiten.* Weinheim: Beltz.

Adresse des Autors:

Dr. Peter Birkel
PH Weingarten
Fach Pädagogische Psychologie
Kirchplatz 2
88250 Weingarten
Fon: 0751-52007
Fax: 0751-52027
E-Mail: birkel@ph-weingarten.de

Manuskripteingang: 14.Januar 2004
Typoskripteingang: 02.Dezember 2004

Anhang:

A) Klassenarbeit Klasse 4 (Im Jahr 2000 wurde noch mit DM bezahlt!)

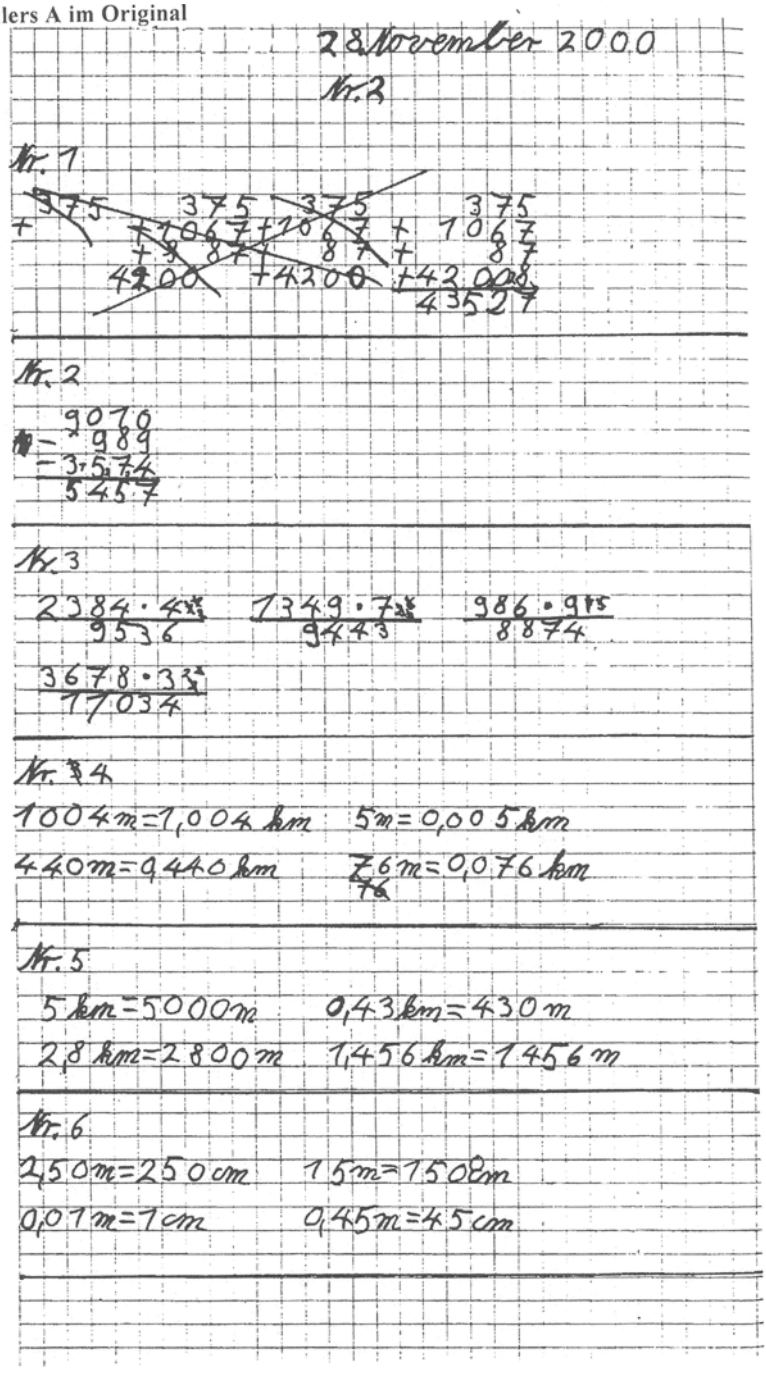
1) Schreibe untereinander und addiere:
   375 + 1067 + 87 + 42008
2) Schreibe untereinander und subtrahiere:
   9010 – 989 – 3574
3) Multipliziere:
   2384 • 4 1349 • 7 986 • 9 3678 • 3
4) Schreibe als Kilometer (Kommazahlen):
   1004 m = 5 m = 440 m = 76 m =
5) Schreibe als Meter:
   5 km = 0,43 km = 2,8 km = 1,456 km =
6) Schreibe als Zentimeter:
   2,50 m = 15 m = 0,01 m = 0,45 m =
7) Schreibe als Meter (Kommazahlen):
   650 cm = 7 cm = 4375 cm = 34 cm =
8) In einer Saftkellerei werden täglich 2800 l Apfelsaft in Literflaschen, 640 l Saft in ¼ l-Flaschen, 800 l Saft in ½ l-Flaschen abgefüllt.
   Wie viele Flaschen werden insgesamt abgefüllt?
9) Der Kilometerzähler von Elkes Fahrrad zeigte vor dem Ausflug 239,7 km.
   a) Nach dem Ausflug stand der Zähler auf 285,3 km.
   b) Abends fuhr Elke noch 7600 m.
10) Kaufmann Kunz bestellte 468 m Kleiderstoff, 1 m zu 9 DM.
11) Im Kaufhaus werden 4 Geschirrspüler zu je 1296 DM und 6 Waschautomaten zu je 899 DM verkauft.
12) Frau Moser hat für 20 Knäuel Wolle 110 DM bezahlt. Sie braucht für den Pullover aber nur 16 Knäuel. Die restlichen gibt sie zurück.
    Wie viel Geld erhält sie dafür?

Viel Glück!

Liebe Kollegin! Lieber Kollege!
Bitte beurteilen Sie die beigefügten Mathematikarbeiten und teilen Sie uns mit, wie Sie zu Ihren Noten gekommen sind. Welche Schritte haben Sie unternommen? Warum haben Sie welche Zensuren erteilt? Notenspiegel?
Bitte geben Sie jedem Schüler individuell Rückmeldung!
Herzlichen Dank für Ihre Bereitschaft zur Mitarbeit!

Dr. Peter Birkel, PH Weingarten

B) Arbeit des Schülers A im Original



Nr. 7

650 cm = 6,50 m 7 cm = 0,07 m

4375 cm = 43,75 m 34 cm = 0,34 m

Nr. 8

R: 2800 l + 1640 l = 3440

3440 : 4 = 860 R 0

3440 : 4 = 800 R 40

800 l : 2 = 400

A: Es bleiben 850 R 2 Flaschen übrig.

Nr. 9

F: Wieviel km ist sie gefahren?

239,7 km

285,3 km

1954,4 km 239,7 km + 566 km = 285,3 km

A: Sie ist noch 566 km gefahren.

R: 285,3 km − 7600 km = 5253 km

A: Sie ist dann 5253 km gefahren.

F: Wieviel DM muss sie bezahlen?

R: 468 : 9 = 46,2 DM 4602 DM

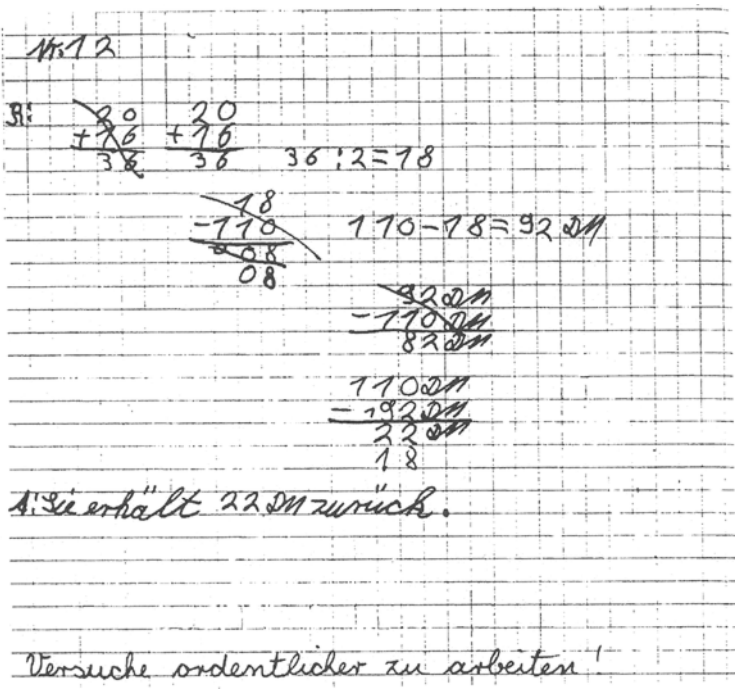
A: Es kostet 4602 DM

Nr. 11

F: Wieviel wird verkauft.

R: 7236 DM · 4/7 = 5184 DM 899 DM · 6/7 = 5394 DM

A: Es wird 5184 DM und 5394 DM.


Nr. 12
A: 20 + 16 = 36
20 + 16 = 36
36 : 2 = 18
110 − 18 = 92 DM
92 DM − 110 DM = 82 DM
110 DM − 92 DM = 18 DM
A: Sie erhält 22 DM zurück.
Versuche ordentlicher zu arbeiten!