Die graue Maus

Eine empirische Untersuchung, ob eine Skalenbewertung bei der Bewertung von Deutschreferaten zu höherer Übereinstimmung zwischen Beurteilern führt

Universität Utrecht
Deutsche Sprache und Kultur: Bildung und Kommunikation
Masterarbeit

Begleitung:	Prof. dr. H. van den Bergh
	D. Abitzsch, MA
Studierende:	T.S. Witte, BEd
Matrikelnummer:	3040372
Datum:	28-10-2020

Inhaltsverzeichnis

Samenvatting

In het voortgezet onderwijs leidt de beoordeling van spreekvaardigheid Duits niet zelden tot een onwenselijke situatie: leerlingen krijgen vaak van verschillende leerkrachten Duits voor een gelijkwaardige mondelinge prestatie een verschillend cijfer. Waar de leerling bij de ene leraar een 7 krijgt, krijgt diezelfde leerling bij een andere leraar een 5. In dit empirische onderzoek is in exploratieve zin onderzocht of het mogelijk is om via schaalbeoordeling de beoordelaarsovereenstemming onder leerkrachten Duits te vergroten.
Aan het empirisch onderzoek hebben 12 leraren Duits meegedaan van uiteenlopende leeftijd (gemiddeld 50,7 jaar) en met uiteenlopende leservaring (2 tot 49 jaar, gemiddeld 22,9 jaar). Deze 12 leraren moesten (op audio opgenomen) spreekbeurten van 25 4vwo-leerlingen beoordelen in twee rondes: in de eerste ronde beoordeelden zij die spreekbeurten globaal, dus op basis van hun eigen niet-geëxpliciteerde criteria. In de tweede ronde die enkele weken later plaatsvond, beoordeelden ze diezelfde spreekbeurten aan de hand van een zogenaamde ‘grijze muis’ (de spreekbeurt die het cijfer ‘5’, dus het gemiddelde van het Nederlandse cijfersysteem, representeerde).
In de tweede beoordelingsronde werd een specifieke vorm van schaalbeoordeling toegepast, namelijk *magnitude estimation* ontwikkeld door Stevens (1975). Daarbij moesten de leraren telkens inschatten of ze een concrete spreekbeurt slechter of beter vonden dan de grijze muis; tevens moesten zij aangeven hoeveel keer slechter of beter zij de spreekbeurt vonden.
Verondersteld werd dat de tweede ronde met schaalbeoordeling een hogere beoordelaarsovereenstemming zou opleveren doordat de leraren immers een vast referentiepunt hadden in de vorm van de grijze muis; daar konden ze zich telkens aan vast klampen, iets wat bij de eerste beoordelingsronde niet het geval was.
Uit de resultaten bleek echter dat de globale beoordeling juist tot een grotere overeenstemming leidde vergeleken met de schaalbeoordeling. Het empirisch onderzoek resulteerde echter ook in een positieve uitkomst: de globale- en grijze-muis-beoordelingen bleken qua betekenis niet van elkaar te onderscheiden, zodat geconcludeerd kon worden dat de leraren bij beide

beoordelingen, hoe procedureel verschillend ook, precies dezelfde criteria toepasten.
Als meest plausibele verklaring voor het falen van de hypothese werd verondersteld dat de groep leraren een vrij hoog gemiddeld aantal jaar leservaring had en als gevolg daarvan in al die jaren voor zichzelf een min of meer vast referentiekader had opgebouwd. In de eerste beoordelingsronde konden ze terugvallen op dat vertrouwde kader, maar in de tweede beoordelingsronde waarbij ze op een geheel nieuwe manier moesten beoordelen, ontbeerden ze juist die ervaring. Dat zou betekenen dat met name jonge leerkrachten met weinig leservaring baat zouden hebben bij schaalbeoordeling.

Zusammenfassung

In der weiterführenden Schule führt die Bewertung der mündlichen Leistungen im Fach Deutsch häufig zu unerwünschten Situationen: Die Schüler/innen werden von unterschiedlichen Lehrkräften für eine vergleichbare mündliche Leistung unterschiedlich zensiert. Während eine Lehrkraft eine Leistung mit einer 2 benotet, wird die gleiche Leistung bei einer anderen Lehrkraft mit einer 6 benotet. In dieser empirischen Studie wurde in explorativem Sinne untersucht, ob es möglich ist, mit einer Skalenbewertung die Übereinstimmung in der Bewertung bei den Lehrkräften für Deutsch zu erhöhen.

12 Lehrkräfte für Deutsch haben an dieser empirischen Studie teilgenommen. Das Alter der Lehrkräfte (durchschnittlich 50,7 Jahre) sowie die Unterrichtserfahrung (2 bis 49 Jahre, durchschnittlich 22,9) variierten stark. Die 12 Lehrkräfte bekamen die Aufgabe, (Audio-) Aufnahmen von Referaten von 25 Zehntklässlern eines Gymnasiums in zwei Runden zu bewerten: In der ersten Runde wurden die Referate ganzheitlich, das heißt auf der Grundlage ihrer eigenen nicht-expliziten Kriterien, beurteilt. In der zweiten Runde, die einige Wochen später stattfand, bewerteten sie dieselben Referate mit der sogenannten „grauen Maus“ (das Referat, das mit der Note 5 und damit dem Durchschnitt des niederländischen Notensystems bewertet wurde).

In der zweiten Bewertungsrunde wurde eine bestimmte Form der Skalenbewertung angewendet, die von Stevens (1975) entwickelte *Potenzfunktion*. Dabei haben die Lehrkräfte jedes Mal ein bestimmtes Referat entweder schlechter oder besser als die „graue Maus“ eingestuft; zudem haben sie angegeben, wie viele Male schlechter bzw. besser das Referat war. Es wurde angenommen, dass in der zweiten Runde mit der Skalenbewertung eine höhere Übereinstimmung in der Bewertung erzielt werden würde, da die Lehrkräfte einen festen Bezug in Form der „grauen Maus“, an der sie sich orientieren konnten, zur Verfügung hatten, dies war in der ersten Bewertungsrunde nicht der Fall.

Aus den Ergebnissen folgte allerdings, dass im Vergleich zu der Skalenbewertung die ganzheitliche Bewertung zu einer höheren Übereinstimmung führte. Die empirische Studie führte allerdings auch zu

einem positiven Ergebnis: Sowohl die ganzheitlichen Bewertungen als auch die „graue Maus“-Bewertungen unterscheiden sich, von der Bedeutung her, nicht voneinander, sodass geschlussfolgert werden kann, dass die Lehrkräfte bei beiden Bewertungen, obwohl methodisch abweichend, genau dieselben Kriterien angewendet haben.

Eine plausible Erklärung für das Scheitern der Hypothese wird darin gesehen, dass die Lehrkräfte durchschnittlich sehr viele Jahre an Unterrichtserfahrung mitbrachten. Die Folge davon ist, dass sie im Laufe der Jahre einen mehr oder weniger festen Referenzrahmen für sich erstellt haben. In der ersten Bewertungsrunde konnten sie auf den vertrauten Referenzrahmen zurückgreifen, in der zweiten Bewertungsrunde jedoch musste ihre Beurteilung auf eine ganz neue Art und Weise durchgeführt werden, hier fehlte ihnen ihre Erfahrung. Das würde bedeuten, dass vor allem junge Lehrkräfte mit wenig Unterrichtserfahrung von der Skalenbewertung profitieren könnten.

I Hintergrund

Als ich 2012 als frischgebackener und engagierter Deutschlehrer anfing, lagen mir zwei Kompetenzen am Herzen: der mündliche Ausdruck und die Konversationskompetenz, d. h. das Referat und der Dialog. Die Lehrerausbildung (in den Niederlanden hbo: Fachhochschule) stellte damals diese Fähigkeiten in den Mittelpunkt. Man hatte dabei eine Interpretation im Sinne des „Gemeinsamen europäischen Referenzrahmens für Sprachen“ (GER) sowie die „Kann-Beschreibungen“ vor Auge; selbstverständlich richtete ich meinen Unterricht darauf aus.
Daher legte ich in meinem Deutschunterricht immer großen Wert auf die mündliche Ausdrucksweise; die Schüler/innen erhielten die Aufgabe, sich vorzustellen, etwas über ihre Familie oder ihr Haustier, über Hobbys, Musiktitel, einen Zeitungsartikel, ein Buch und dergleichen zu erzählen. Natürlich musste ich den mündlichen Ausdruck benoten, das war normalerweise ein Befriedigend - vorausgesetzt, der/die Schüler/in hatte sich angestrengt und der Kern des mündlichen Referats kam bei mir an.
Im Laufe der Zeit stellte ich jedoch fest, dass meine Kollegen (in diesem Fall Lehrkräfte für Deutsch) ihren Schülern viel schlechtere bzw. viel bessere Noten für die von ihnen bewerteten Referate gaben. Auf Anfrage stellte sich heraus, das viele Lehrkräfte, übrigens auch ich, bei der Bewertung des mündlichen Ausdrucks nach eigenen Auskünften „einfach etwas gemacht haben“: Wenn das Referat nach ihrem subjektiven Eindruck und Urteil in Ordnung war, dann bekam es eine 7. Andere Lehrkräfte dagegen urteilten auf der Basis expliziter Bewertungskriterien, diese Kriterien waren: korrekte Aussprache und Konjugation bestimmter Verben, Sprachgewandtheit usw.

Kurz gesagt, in meiner neunjährigen Karriere als Deutschlehrer an vier unterschiedlichen Schulen stieß ich jedes Mal auf das gleiche Problem: Verschiedene Lehrkräfte zensierten die mündlichen Leistungen ihrer Schüler/innen im Fach Deutsch auf unterschiedliche Weise. Die Folge war, dass der/die Schüler/in für eine objektiv gesehen gleichwertige Leistung bei zwei verschiedenen Lehrkräften einmal eine 7 und das andere Mal eine 5 bekam.

2018 leitete ich meine erste Abiturklasse und war damit auch verantwortlich für die mündliche Abiturprüfung. Damit die Bewertung objektiver wurde, wurden (und werden) zwei Lehrkräfte zur Bewertung eingesetzt. Aus dem GER haben wir beide die unserer Meinung nach wichtigsten Kategorien ausgewählt und die Leistungen der Schüler/innen anhand dieser Kategorien zensiert. Und doch lief es oft folgendermaßen ab: „Was denkst du?“ „Ich finde es in Ordnung, sollen wir eine 7 geben?“ „Ja, ist gut.“ Ich persönlich fand es beispielsweise sehr schwierig, ob ich nun 1 oder doch 2 Punkte für eine bestimmte Kategorie aus dem GER geben sollte...das zu beurteilen ist und bleibt doch immer auch subjektiv?

1.1 Zielsetzung

Das beschriebene Problem - Lehrkräfte sind sich nicht einig über die Note, die sie für dieselbe Leistung vergeben - ist ein großes Problem, denn es untergräbt das von De Groot (1966) erwähnte „Ideal der Chancengleichheit“. Ein/e Schüler/in muss in einer Prüfung davon ausgehen können, dass die abgelegte *Leistung* so fair und gerecht wie möglich beurteilt wird: Die gleiche Leistung wird gleich benotet. Aber in der Praxis und aus der umfangreichen Forschungsliteratur geht hervor, dass dies keineswegs immer der Fall ist. Ein extremes Beispiel dafür stammt aus einer groß angelegten deutschen, empirischen Untersuchung über die Bewertung von Aufsätzen (1113 deutsche Lehrkräfte für Deutsch nahmen an der Studie teil, in der insgesamt 617 Aufsätze zensiert wurden). Das Ergebnis (laut *Die Zeit*, Nr. 49, 4. 12. 1970) lautete: „Der eine gibt „1” und der andere „6” – und das für genau den gleichen Aufsatz! (Meuffels, 1994, S. 42).

Hier wird deutlich, dass die oben genannte niedrige Interrater-Reliabilität (von jetzt an: IRR) unerwünscht ist. In dieser Arbeit möchte ich anhand einer von mir durchgeführten empirischen Studie herausfinden, ob die Möglichkeit besteht, über eine sogenannte Skalenbewertung die niedrige IRR zu eliminieren und wesentlich zu erhöhen.

Lehrkräfte haben bei der Zensierung oft keinen klaren Bezugsrahmen und geben deshalb ‚einfach aufs Geratewohl' eine Note - das ist meistens eine „unverfängliche" 6 oder 7 (De Groot, 1966). Bestenfalls steht den Lehrkräften ein Standard zur Verfügung, den sie selbst wichtig finden. Auf dessen Grundlage beurteilen sie dann, ob eine Leistung ausreichend oder mangelhaft ist. Was wäre, wenn wir allen Lehrkräften exakt die gleiche Vergleichsbasis geben, genau den gleichen Bezugsrahmen? Was wäre, wenn alle Lehrkräfte den mündlichen Ausdruck ihrer Schüler/innen anhand eines exakt gleichen Ankers zensieren würden: Ein konkretes Beispiel für ein Referat, das von einer Fachjury weder schlecht noch gut bewertet würde, also ein konkretes Beispiel, das zu Recht als die „graue Maus" bezeichnet werden kann? *Müssten* dann die Zensuren nicht besser übereinstimmen als das normalerweise der Fall ist?

1.2 Aufbau

Diese Arbeit ist folgendermaßen aufgebaut. In Kapitel II wird der theoretische Hintergrund der empirischen Untersuchung beschrieben. Zunächst werden in diesem Kapitel verschiedene Störfaktoren, die für die mangelnde Übereinstimmung zwischen verschiedenen Beurteilern verantwortlich sein können, kurz aufgelistet und erläutert. Darüber hinaus wird in dem Kapitel die Frage untersucht, welcher der genannten Störfaktoren höchstwahrscheinlich durch die Skalenbewertung beseitigt werden kann und weshalb erwartet wird, dass eine Skalenbewertung einen Zuverlässigkeitsgewinn (das heißt mehr Übereinstimmung zwischen den Beurteilern sowie eine höhere IRR) mit sich bringt. Die Überlegungen in diesem Kapitel führen zu einer konkreten Fragestellung bzw. Hypothese, die auf der Grundlage von gesammeltem Zensur-Material aus Bewertungen von 25 Deutschreferaten von Zehntklässlern eines Gymnasiums empirisch überprüft werden soll.

In Kapitel III werden unterschiedliche Formen der Zuverlässigkeit erörtert. Es wird auch die Zuverlässigkeitsform spezifiziert, die sich in der vorliegenden empirischen Studie über eine Skalenbewertung (erwartungsgemäß) verbessern sollte und wie sich dies anschließend an den Zensuren zeigen

sollte. In dem Kapitel wird auch diskutiert, dass die Skalenbewertung (laut Forschungsliteratur) viele Varianten aufweist.

Kapitel IV enthält eine detaillierte Beschreibung der Aufstellung und Durchführung der empirischen Studie. Es wird unter anderem begründet, warum in dieser Studie nur ein einzelner Anker, der den „grauen Durchschnitt“ repräsentiert, ausgewählt wurde; weshalb der grauen Maus die Maßzahl 100 zugewiesen wurde und warum es den Beurteilern überlassen bleibt, die Qualität des zu bewerten Produkts in einer Maßzahl in Relation zu dem numerisch festgelegten Anker von 100 auszudrücken, ohne dass hierfür weitere Spezifikationen angegeben werden. Die verschiedenen Schritte, die in der Studie zur Beseitigung störender Faktoren unternommen wurden, werden begründet und diskutiert.

In Kapitel V werden die Ergebnisse der empirischen Untersuchung beschrieben.

In Kapitel VI werden die erzielten Untersuchungsergebnisse diskutiert und es werden einige Vorschläge für mögliche weitere Untersuchungen vorgestellt.

II Theoretischer Hintergrund

Die Zielsetzung dieser Arbeit beruht auf der These, dass bei der Zensierung von Referaten von Lehrkräften für Deutsch im Allgemeinen eine geringere IRR bzw. eine so geringe Zuverlässigkeit erreicht wird, dass die Skalenbewertung effektiv und sinnvoll sein kann. Ist diese These aus empirischer Sicht richtig, und wenn ja, wie könnte diese schlechte Zuverlässigkeit erklärt werden? Weshalb sollte eine Skalenbewertung zu einer höheren Zuverlässigkeit führen?

2.1 Interrater-Reliabilität (Beurteilerübereinstimmung) empirisch untersucht

Seit Beginn des 20. Jahrhunderts wurden in empirischen Studien immer wieder erhebliche Unterschiede bei der Benotung durch Lehrkräfte festgestellt (Van den Ende, 1954; De Groot, 1961; De Groot,1966; Wesdorp,1981; Meuffels, 1994). Diese Unterschiede treten bei Weitem nicht nur in den „weichen Fächern“ wie Fremdsprachen, Geschichte oder Erdkunde auf. 1912/1913 wurde bereits in der Untersuchungsreihe von Starch & Elliot überzeugend dargelegt, dass sogar in einem „harten Fach“ wie Mathematik manchmal große Unterschiede zwischen den Benotungen der Lehrkräfte auftreten. 118 Lehrkräfte für Mathematik sollten genau denselben Test in der entsprechenden Studienreihe nach dem sogenannten „Percentage Grading“-System bewerten, bei dem die Bewertungsskala von 1 bis 100 reicht, mit einer ausreichenden/unzureichenden Punktzahl von 70. Es wurde eine enorme Diskrepanz festgestellt, die von 28 bis zu 92 Punkten reichte! 24 Lehrkräfte benoteten den Test mit weniger als 60 Punkten (d. h. ungenügend) und 20 Lehrkräfte bewerteten die Mathearbeit mit 80 oder mehr Punkten, damit gaben sie also die Note „Gut“ oder „Sehr gut“. Dieses schockierende Ergebnis kann kaum als Zufall bezeichnet werden, wie eine amerikanische Studie von William (1933) zeigte. In seiner Studie sollten 100 Lehrkräfte für Mathematik 50 (Mathematik-)Prüfungen nachschauen und bewerten, wiederum nach dem in Amerika üblichen „Percentage Grading“-System. Im

extremsten Fall wurden sogar Unterschiede von 16 bis 96 Punkten erreicht (Meuffels, 1994, S.47-48).
In den darauffolgenden Jahren gab es einen stetigen Strom von in- und ausländischen Publikationen, in denen der Mangel an Einstimmigkeit zwischen den Benotungen - unabhängig von dem Unterrichtsfach - beschrieben wurde (Van den Ende, 1954; De Groot, 1961; De Groot,1966; Wesdorp,1981; Meuffels, 1994). Van den Ende schlussfolgert 1954 nach einer gründlichen Analyse der empirischen Daten, dass bei der Zensierung eher eine „zügellose Subjektivität“ herrscht als eine fundierte, sachliche Bewertung. Eine ähnliche Beschwerde stammt von De Groot (1966) in seinem kontroversen *Vijven en zessen (Sechsen und Fünfen)*, in dem er behauptet, dass die Benotung eher auf den Beurteiler selbst als auf die Leistung zutrifft. Angesichts der zahlreichen negativen Forschungsergebnisse, aufgrund derer die Kompetenz der Lehrkräfte als bewertende Instanz als „gering“, „ungenügend“, „schlecht“, „abwesend“, „nicht gut“ oder „ziemlich bedauerlich“ bezeichnet wird (siehe u. a. Hofmann & Verbeek 1977: S.143; Kreeft, Luyten & Schreuder 1978: S.102-103; Meuffels 1978: S.57; Rijlaarsdam & Blok 1981: S.753-754), gibt es wenig bis gar keine Zweifel daran, dass die These in der Zielsetzung dieser Arbeit („bei der Bewertung von Referaten durch Lehrkräfte für Deutsch besteht im Allgemeinen (...) eine niedrigere IRR“) zu Recht besteht. Diese Arbeit basiert daher voll und ganz auf diesem Ausgangspunkt.

2.2. Störfaktoren

Angesichts der relativ ähnlichen empirischen Ergebnisse für die IRR bei der Benotung durch Lehrkräfte in der oben aufgeführten Übersicht stellt sich die Frage: Wie kann ein solcher (in der Regel erheblicher) Mangel an Zuverlässigkeit bei der Benotung durch Lehrkräfte erklärt werden? Welche Faktoren könnten dafür verantwortlich sein?
In einer theoretischen Analyse der Bewertungsproblematik summieren De Groot (1961), Wesdorp (1981), Schoonen (1991) und Meuffels (1994) einige Störfaktoren, die zu einer niedrigen IRR führen (können):

- Bedeutungseffekt (Beurteiler achten auf unterschiedliche Aspekte)
- Halo-Effekt
- Kontamination im weiteren Sinne
- Kontamination im engeren Sinne
- Sequenzeffekt
- Normverschiebung
- Persönlicher Vergleich

Oben genannte Faktoren werden im Folgenden kurz anhand eines hypothetischen Beispiels erklärt: Zwei Lehrkräfte für Deutsch, die in einer Schulprüfung die Deutschreferate einer Reihe von niederländischen Schüler/innen beurteilen sollen - und die Qualität dieser Referate in einer Zensur ausdrücken sollen.

Die Benotung durch die beiden Lehrkräfte kann unterschiedlich sein, weil die Bewertungsaufgabe von beiden unterschiedlich aufgefasst wird, bzw.: Sie weisen der Bewertungsaufgabe eine andere Bedeutung zu. Denn was genau ist nun ein guter mündlicher Ausdruck im Deutschen und woraus sollte dieser *konkret* bestehen? Der eine achtet auf die Sprachgewandtheit, der andere auf einen grammatikalisch korrekten Satzaufbau, wieder ein anderer achtet darauf, ob klar und deutlich wird, was der/die Schüler/in sagen möchte usw. Persönliche Unterschiede in der Interpretation der Bewertungsaufgabe, genannt der *Bedeutungseffekt*, sind oft für eine niedrige IRR verantwortlich.

Außerdem kann die Benotung durch zwei Lehrkräfte unterschiedlich sein, weil die eine Lehrkraft ihre „eigenen“ Schüler/innen kennt und deshalb über besondere Vorkenntnisse über sie verfügt, Vorkenntnisse, die der zweite Beurteiler nicht besitzt. Wenn die Rednerin beispielsweise ein braves und aufmerksames Mädchen in der Klasse ist, das im Durchschnitt immer gut mitmacht, wird sie auch bei dem Referat gut abschneiden - zumindest bei ihrer eigenen Lehrkraft. Aber ein nervender, lauter Schüler, der hinten in der Klasse sitzt und immer den Unterricht stört, wird es sehr viel schwerer bei seiner eigenen Lehrkraft haben. Wenn diese spezifischen Kenntnisse eines/r zu beurteilenden Schülers/Schülerin die Bewertung der mündlichen

Deutschkenntnisse - ob nun bewusst oder unbewusst - beeinflussen, spricht man von dem *Halo-Effekt*: Die störende „Ausstrahlung“ auffälliger Eigenschaften, die mit den Merkmalen, die eigentlich beurteilt werden sollten, nichts zu tun haben.

Die Note für die abgelegte Leistung muss im sogenannten „guten gegenseitigen Einverständnis“ zwischen den beiden Lehrkräften für Deutsch zustande kommen. Aber was bedeutet *gut* in dieser Situation? Die eine Lehrkraft könnte sich in ihrer Bewertung unbewusst beeinflussen und beeindrucken lassen, beispielsweise durch den autoritären Stil des anderen Beurteilers, der um jeden Preis seine eigene, für heilig erklärte Note durchsetzen möchte. Bewertungen sollten selbstverständlich unabhängig voneinander zustande kommen oder anders ausgedrückt: Eine Lehrkraft darf sich nicht von den Kommentaren und / oder Bewertungen der anderen Lehrkraft beeinflussen lassen. Wenn dies der Fall ist, dann sind die Bewertungen *kontaminiert im weiteren Sinne*. Es ist übrigens logisch, dass, im Gegensatz zu dem Bedeutungs- und dem Halo-Effekt, eine Kontamination im weiteren Sinne in der Regel zu einer Erhöhung der Interrater-Reliabilität führt; schließlich lässt sich die eine Lehrkraft von der anderen beeinflussen.

Eine Lehrkraft, die in einer Prüfung die mündlichen Leistungen ihrer eigenen Schüler/innen beurteilen muss, hat im Grunde ein persönliches Interesse an den Ergebnissen ihrer Bewertung: Welche Lehrkraft würde die mündlichen Leistungen von allen ihren Schüler/innen (und mit „allen“ sind hier wirklich alle gemeint, ohne Ausnahmen) mit einem „Ungenügend“ benoten - selbst wenn die Leistungen aus objektiver Sicht wirklich alle unterdurchschnittlich wären? Denn dies würde den definitiven Todesstoß für den eigenen Status als Lehrkraft bedeuten: Diese Person hätte als Lehrkraft total versagt. Diese Freiheit im Bewertungssystem ermöglicht es den beiden Lehrkräften allerdings, bewusst oder unbewusst, andere Interessen und Ziele in ihre Bewertung einfließen zu lassen als die einer streng unparteiischen Bewertung. Sollten ihre Bewertungen durch solche Interessen eingefärbt sein, sind sie *kontaminiert im engeren Sinne*. Wenn die Interessen der beiden

Lehrkräfte in unserem Beispiel voneinander abweichen, kommt dies der IRR in der Regel nicht zugute.

Auch der sogenannte *Sequenzeffekt* (eine Bewertung ist nicht unabhängig von der vorhergehenden) kann die Benotung unbeabsichtigt beeinflussen: Wenn die Lehrkräfte die ersten vier Referate sehr schlecht bewertet haben, werden sie bei dem darauffolgenden mäßigen Referat erleichtert aufatmen und diesem Referat, zu Unrecht, eine zu hohe Note geben. Sollte sich allerdings das gleiche mäßige Referat an vier sehr gute Referate anschließen, wird die Note sehr viel schlechter ausfallen. Der Einfluss dieses Störfaktors auf die IRR ist unterschiedlich: Wenn die zwei Lehrkräfte die Referate in genau derselben Reihenfolge bewerten, wird die Übereinstimmung wahrscheinlich zunehmen. Sollten sie dagegen die Referate in einer vollkommen anderen Reihenfolge bewerten müssen, nimmt die Übereinstimmung erheblich ab.

Das Auftreten eines Sequenzeffekts in dem oben geschilderten Beispiel zeigt, dass die Normen der beiden Beurteiler nicht feststehen, sondern „je nach den Umständen“ variieren und sich verschieben. Angenommen, das Deutschreferat eines/r Schülers/Schülerin muss sich an vier anderen, hervorragenden Referaten messen lassen, dann wird das Referat wahrscheinlich nicht gut abschneiden - sollte es allerdings mit vier schlechteren Referaten konkurrieren, dann würde die Note besser ausfallen. Normen sind selten absolut und auch nicht stabil. Bewertungen passen sich u. a. an das Niveau einer Gruppe von Produkten an, die als Ganzes bewertet werden. Was die Wirkung dieser sogenannten *Normverschiebung* auf die IRR betrifft: In der überwiegenden Mehrzahl der Fälle wird sie negativ ausfallen.

Selbst wenn die oben genannten Faktoren in dem hier diskutierten hypothetischen Fall der beiden Lehrkräfte für Deutsch, die in einer Prüfung die Referate der Prüfungskandidaten bewerten müssen, nicht auftreten würden, besteht noch stets ein beträchtliches Risiko, dass die beiden nicht einstimmig urteilen werden. Jeder Einzelne neigt schließlich dazu, auf eine für ihn charakteristische Weise zu urteilen: Der eine ist sehr streng, der andere

flexibel, beim nächsten regnet es Fünfen und Sechsen usw. Diese Art der Bewertungsgewohnheiten ist in der Literatur bekannt unter dem Begriff *persönlicher Vergleich.* Die Auswirkung auf die IRR ist negativ.

2.3 Maßnahmen zur Beseitigung von Störfaktoren

Angesichts dieses Dilemmas in der schulischen Benotung ist es nicht verwunderlich, dass seit Langem nach wirksamen Maßnahmen, die die Störfaktoren eliminieren können, geforscht wird und damit - hoffentlich - die Übereinstimmung zwischen den Lehrkräften steigern werden. Ende der sechziger bis Anfang der siebziger Jahre wurde in den Niederlanden (zum Teil aufgrund des Buches *Vijven en Zessen* von De Groot, 1966) in der Sekundarstufe I eine Maßnahme eingeführt, die den Störfaktoren ein sofortiges Ende bereitete: der objektive Unterrichtstest. Der objektive Unterrichtstest wurde (und wird heutzutage häufig noch) als die Lösung schlechthin für die niedrige IRR angesehen: Der Mensch (bzw. die Lehrkraft) wird beim Nachschauen, Korrigieren und Benoten von Prüfungen, die die Schüler/innen ablegen, komplett abgeschafft und von einer Maschine ersetzt: dem Computer. Wenn ein Computer die Aufgaben korrigiert, ist es natürlich ausgeschlossen, dass die oben genannten Störfaktoren das Ergebnis beeinflussen. Was die Bewertung der mündlichen Deutschkenntnisse der Schüler/innen anbelangt, ist es derzeit noch nicht möglich, diese Bewertung streng objektiv vorzunehmen: Wir sind also vorerst noch auf den Menschen angewiesen.

Im Laufe der Zeit wurden viele spezifische Vorsorgemaßnahmen vorgeschlagen, um einen oder mehrere Störfaktoren auszuschalten und somit eine zuverlässigere Bewertung zu gewährleisten. Im Folgenden werden einige der Maßnahmen kurz erläutert.

Es werden analytische Schemata vorgeschlagen, um den *Bedeutungseffekt* auszuschalten. Referate können auf globale / ganzheitliche oder auf analytische Weise beurteilt werden. Im Gegensatz zu der globalen Bewertung, bei der der Beurteiler von seinen eigenen, nicht-expliziten

Normen ausgeht und die Referate nach dem Gesamteindruck bewertet (was zu einer Note führt), wird bei der analytischen Bewertung die „Gesamtqualität“ in unterschiedliche, mehr oder weniger unabhängige Teilaspekte unterteilt (z. B. Satzaufbau, Wortschatz, Sprachgewandtheit, Aussprache). Jedes Referat muss nach all diesen verschiedenen Aspekten separat beurteilt werden; die analytische Abschlussbewertung besteht normalerweise aus dem Notendurchschnitt der verschiedenen Teilaspekte. Jedoch sind die Ergebnisse der empirischen Untersuchung, in der beide Bewertungsmethoden in Bezug auf die Zuverlässigkeit miteinander verglichen werden, anders als erwartet (nämlich: analytische Schemata führen zu einer höheren IRR) „enttäuschend“, wie Wesdorp in seinem umfassenden Überblick einer relevanten Studie zeigte (Wesdorp 1981: S.57-58) (siehe auch Meuffels, 1994 und Van den Bergh und Meuffels, 2000). Eine Erklärung, warum analytische Bewertungsmodelle nicht funktionieren, gibt Meuffels (1994): Die Ursache liegt seiner Meinung nach darin, dass die Angaben und Beschreibungen der analytischen Bewertungskategorien oft unklar und mehrdeutig sind (beispielsweise Stil, Originalität, Aufbau usw.). Er behauptet, dass der allgemeine Gesamteindruck dominiert: Wenn der Eindruck positiv ist, werden die Merkmale des analytischen Schemas auch positiv bewertet. Umgekehrt gilt das Gleiche, was bedeutet, dass eine globale und eine analytische Bewertung tatsächlich die gleichen Qualitäten widerspiegeln - und daher gleichermaßen zuverlässig sind.

Der Halo-Effekt kann ausgeschaltet werden, indem eine Jury hinzugezogen wird, deren Mitglieder (außer dem/r Klassenlehrer/in) keine Vorkenntnisse über den/die betreffende/n Schüler/in haben - was in der täglichen Schulpraxis allerdings schwer zu organisieren ist.

Der Kontaminationseffekt im weiteren Sinne kann bekämpft werden, indem beide Lehrkräfte *unabhängig voneinander* die Referate benoten; nach der Bewertung können beide Zensuren miteinander verglichen und diskutiert werden. Auch diese Option ist in der Praxis kaum durchführbar.

Der Kontaminationseffekt im engeren Sinne kann bekämpft werden, indem zwei Lehrkräfte eingesetzt werden, die keinerlei Interesse am Ergebnis einer Prüfung haben. Das würde bedeuten, dass der/die Klassenlehrer/in nicht als Prüfer/in auftreten darf: erneut eine kaum durchführbare Option.

Dem Sequenzeffekt kann entgegengewirkt werden, indem die Referate erneut bewertet werden, dann allerdings in einer anderen, willkürlichen Reihenfolge. Das würde bedeuten, dass die Referate der Schüler/innen aufgezeichnet werden müssten. Sollte jedoch das Aufnahmegerät eine/n der betroffenen Schüler/innen beeinflussen, z. B. durch Angst, dann ist auch diese Option nicht durchführbar.

Der Normverschiebung kann entgegengewirkt werden, indem Lehrkräfte einen festen Standard zur Hand bekommen, konkrete Beispielreferate (Anker), die bereits benotet sind. Die Aufgabe des Beurteilers besteht dann darin, jedem zu bewertenden Produkt eine Note in Bezug auf einen Referenzpunkt (einen Anker) auf der angegebenen Skala zu geben. Empirische Untersuchungen zu dieser Art der Skalenbewertung, (in der insbesondere die schriftlichen Leistungen Gegenstand der Untersuchungen sind), zeigen positive Ergebnisse in Bezug auf die Zuverlässigkeit bzw. Übereinstimmung (Wesdorp, 1981; Schoonen, 1991; Pollmann, Prenger en De Glopper, 2012; Pullens, 2012). Da die Skalenbewertung nicht nur einen positiven Effekt auf die Normverschiebung ausübt, sondern auch auf andere Störfaktoren, wird im folgenden Kapitel auf diese vielversprechende Bewertungsmethode eingegangen.

Zusammenfassend lässt sich zu den oben genannten, knapp diskutierten Maßnahmen zur Gewährleistung der Bewertungsobjektivität sagen, dass die (mit Ausnahme der Skalenbewertung) prinzipiell logischen, d. h. offensichtlichen „Notfallmaßnahmen“, aus empirischer Sicht im schwer handhabbaren Schulalltag leider nicht sehr effektiv zu sein scheinen. Kurz gesagt, die Wirksamkeit dieser Maßnahmen ist fragwürdig; die vielleicht beste Lösung besteht darin, die Beurteiler (im Voraus) zu informieren und sie auf

die Existenz von Störfaktoren aufmerksam zu machen, von denen sie möglicherweise nichts wussten.

2.4 Skalenbewertung und Störfaktoren

Der empirische Erfolg der Skalenbewertung, insbesondere bei der Bewertung der schriftlichen Leistungen (Wesdorp, 1981; De Glopper, 1989; Schoonen, 1991; Pollmann et al., 2012; Pullens, 2012; Elving & Van den Bergh, 2015), lässt sich dadurch erklären, dass diese Bewertungsmethode - zumindest theoretisch - gleichzeitig fünf objektivitätsbedrohende Faktoren reduzieren kann: den Bedeutungseffekt , den Halo-Effekt, den Sequenzeffekt, die Normverschiebung und den persönlichen Vergleich (Wesdorp, 1981; Van Schooten, 1988; Pollmann et al., 2012)

Wir versuchen dies zu erklären, wobei unser Ausgangspunkt darin besteht, dass der Beurteiler die schriftlichen Leistungen der Schüler/innen auf einer Skala mit fünf Ankern, die jeweils eine Note (6,5,4,3,2) besitzen, bewerten soll. Die Anker bestehen aus konkreten, kurzen Aufsätzen, die sich (selbstverständlich) qualitativ unterscheiden und die alle benotet wurden (die benoteten Anker wurden mithilfe empirischer Voruntersuchungen von einer Jury aufgestellt, die diese Aufsätze im Großen und Ganzen einstimmig beurteilt hat). Die Aufgabe des Beurteilers besteht dann darin, jedem zu bewertenden Produkt eine Note in Bezug auf einen Referenzpunkt (einen Anker) auf der angegebenen Skala zu geben.

Mithilfe der Skala wird dem Bedeutungseffekt entgegengewirkt: Jeder Beurteiler bekommt eine „operative Definition“ der Begriffe „sehr schlechter“, „schlechter“, „mäßiger“, „ausreichender“ oder „guter“ Aufsatz; er kann folglich alle Aufsätze an denselben Bezugspunkt bzw. denselben Standard koppeln. Den Beurteilern wird also kein Spielraum für eigene Interpretationen der „Qualität eines Aufsatzes“ eingeräumt.

Der Halo-Effekt ist ausgeschaltet, weil der Beurteiler die guten bzw. schlechten Eigenschaften eines/einer Schülers/Schülerin nicht mehr bei seiner Bewertung berücksichtigen kann, da ihm eine Skala vorgegeben wird.

Dem Sequenzeffekt wird entgegengewirkt, indem der Beurteiler die Aufsatzqualität nicht mit den vorangegangenen Aufsätzen vergleichen muss (die möglicherweise sehr schlecht oder sehr gut waren), sondern sich wiederum korrekt an dem gleichen Standard orientieren muss: der Skala.

Der gleiche Standard verhindert außerdem, dass bei der Skalenbewertung eine Normverschiebung auftreten kann: Der Standard steht *unveränderlich* fest.

Zum Schluss ein nicht unwesentlicher Aspekt: Der persönliche Vergleich wird bekämpft oder zumindest reduziert. Die Skala verhindert, dass der Beurteiler „abdriftet“ (wenn beispielsweise ausschließlich extreme Zensuren vergeben werden oder nur unverfängliche Fünfen oder Vieren).

Die persönliche Freiheit des Beurteilers wird durch eine Skalenbewertung erheblich eingeschränkt: Er kann nicht mehr frei nach seinen eigenen, individuellen, undifferenzierten Kriterien beurteilen, sondern er muss anhand eines vorgegebenen Bezugsrahmens urteilen, seine eigenen Präferenzen spielen keine Rolle mehr. Dies ist der Kern (und für einige der Preis) der Objektivierung: die Einschränkung der Handlungsfreiheit des Beurteilers. In dieser Hinsicht nähert sich die Skalenbewertung dem objektiven Unterrichtstest, in dem die Freiheit des Beurteilers nicht mehr vorhanden ist, an.

2.5 Die Hypothese

Aus den vorhergehenden Kapiteln ist deutlich geworden, dass von der Skalenbewertung ein Zuverlässigkeitsgewinn zu erwarten ist: Die Übereinstimmung zwischen den Beurteilern wird größer sein, aber größer als was? Aus methodischer Sicht ist es notwendig, die mit der Skalenbewertung

ermittelte IRR mit einer Referenz zu vergleichen. Die Referenz ist in der aktuellen empirischen Studie die IRR, die durch die Bewertung nach der globalen, ganzheitlichen Methode ermittelt wird.

Damit stellen wir folgende Hypothese auf:

Im Vergleich zu der globalen Bewertung führt die Skalenbewertung zu einer höheren Übereinstimmung zwischen den Beurteilern bei der Bewertung von Deutschreferaten.

Aus dieser Hypothese könnte abgeleitet werden, dass das Durchführen einer empirischen Studie einen reinen prüfenden Charakter hat. In den nächsten Kapiteln wird jedoch deutlich, dass bei der Durchführung der Skalenbewertung der Deutschreferate einige (praktische) Entscheidungen getroffen werden mussten, deren exakte Auswirkungen auf die IRR im Voraus unbekannt waren. Auch die derzeitige Corona-Krise beeinflusst die zu treffenden Entscheidungen. Das führt dazu, dass die durchzuführende Studie weniger einen prüfenden als einen explorativen Charakter besitzt.

III Arten der Zuverlässigkeit (Übereinstimmung) und Skalenbewertung

3.1 Drei Formen der Übereinstimmung

In den vorangegangenen Kapiteln wurden regelmäßig die Begriffe „Zuverlässigkeit“, „Zuverlässigkeitsgewinn“ und „Interrater-Reliabilität“ erwähnt. Aber was bedeutet (in der psychometrischen Fachliteratur) eigentlich *„zuverlässig“*?
Im Hinblick auf die Lehrkraft für Deutsch, die die mündliche Leistung der Deutschreferate ihrer Schüler/innen bewerten soll, bildet „Zuverlässigkeit“ einen mehrdimensionalen Begriff: Der Begriff „zuverlässig“ kann sich auf Folgendes beziehen:

(1) die Stabilität eines Beurteilers (vergibt die Lehrkraft für Deutsch dieselben Zensuren für Deutschreferate ihrer Schüler/innen, wenn sie die Referate nach einigen Wochen erneut bewerten muss?),
(2) die Übereinstimmung zwischen verschiedenen Lehrkräften für Deutsch, die die gleichen Deutschreferate der Schüler/innen bewerten,
(3) sind die Leistungen eines/r Schülers/Schülerin unter unterschiedlichen Konditionen (andere Aufgabe, anderes Thema, andere Situation usw.) mehr oder weniger auf dem gleichen Niveau in Bezug auf den mündlichen Ausdruck?

Die unter (1) und (2) erwähnte Art der Zuverlässigkeit wird Raterstabilität bzw. Interrater-Reliabilität (IRR) genannt, die unter (3) beschriebene Zuverlässigkeit ist in der Fachliteratur unter dem Begriff „Leistungszuverlässigkeit“ bekannt.

In dieser empirischen Studie werden (menschliche) Beurteiler als Messinstrumente eingesetzt: Schließlich bewerten sie die Qualität der Deutschreferate. Bei den Beurteilern (in der Regel Lehrkräfte für Deutsch) kann davon ausgegangen werden, dass sie sich für die abschließende Bewertung auf bestimmte Normen und Werte beziehen oder dass sie einen *Grad der objektiven Spezifizierbarkeit* besitzen (De Groot, 1961). Das

bedeutet, dass jeder Beurteiler nicht „einfach irgendetwas sagt“, sondern dass seine/ihre Bewertung - anhand seiner/ihrer eigenen nicht-expliziten Werte - mehr oder weniger *fundiert* ist. Laut Van den Bergh und Meuffels (2000) sind mündliche Fähigkeiten beispielsweise im Fach Deutsch nicht direkt beobachtbar, das trifft auf alle Sprachkenntnisse zu, ob es produktive oder rezeptive Fähigkeiten betrifft -, stattdessen müssen die Kenntnisse aus verbalen Handlungen, Verhaltensweisen und / oder konkreten Ergebnissen davon abgeleitet werden. Aus dieser Perspektive: Aufgrund der möglicherweise abweichenden Normen und Werte des Beurteilers und der nicht direkten Beobachtbarkeit der mündlichen Fähigkeiten wird deutlich, dass eine zuverlässige Bewertung einer mündlichen Deutschprüfung nicht einfach ist.

Anhand des Beispiels aus Kapitel 2.2, in dem zwei Lehrkräfte eine mündliche Deutschprüfung bewerten, werden die oben genannten Ausführungen erläutert. Die Lehrkräfte für Deutsch in diesem Beispiel können selbstverständlich ihren Grad der objektiven Spezifizierbarkeit beweisen, indem sie sowohl *stabil* als auch intersubjektiv *zuverlässig* in ihrem Urteil sind. Wenn die Lehrkraft *stabil* (bzw. konsistent oder vorhersehbar) handeln möchte, dann muss sie den Deutschreferaten die gleiche Zensur geben, wenn sie die Referate ein paar Wochen später erneut bewertet. Wenn zwei Lehrkräfte Zuverlässigkeit zeigen möchten, *muss* die Bewertung der ersten Lehrkraft mit der der zweiten Lehrkraft in Bezug auf das gleiche Referat übereinstimmen. Die Lehrkraft muss also nicht so sehr mit sich selbst übereinstimmen (sie muss nicht konsistent sein), aber sie muss dem Urteil der anderen Lehrkraft zustimmen (Meuffels, 1994). Wenn beide Lehrkräfte voll und ganz miteinander übereinstimmen, dann spricht man von einer perfekten *Interrater-Reliabilität* (IRR). Zusammenfassend: Für ein zuverlässiges Urteil muss die Subjektivität zwischen den Lehrkräften weitgehend beseitigt werden. Es ist auch notwendig, dass das Urteil sehr stabil bleibt, wenn das Bewertungsverfahren wiederholt wird.
Aber trotz der hohen Stabilität und der hohen IRR kann es sein, dass die Bewertung dieser Lehrkräfte - zumindest in gewissen Aspekten - noch nicht zuverlässig ist. Das kann auf die Variabilität in der Leistung zurückgeführt

werden, auch *Leistungszuverlässigkeit* genannt. Ein/e Schüler/in kann in einer Prüfung beispielsweise hervorragende Leistungen zeigen, wenn es um sein/ihr Hobby geht. Angenommen, ein Schüler ist ein Autofanatiker und darf über dieses Thema ein Referat halten. Dann ist zu erwarten, dass seine Leistung viel besser ist (und entsprechend viel besser bewertet wird) als bei einem Referat über ein Thema, das ihn kaum interessiert. Eine Lösung für dieses Problem der Unzuverlässigkeit könnte so aussehen, dass den Schülern/Schülerinnen ein möglichst breites Spektrum von Themen vorgestellt wird, oder dass das Thema überhaupt nicht festgelegt wird.
Die Lehrkraft, die den mündlichen Ausdruck im Fach Deutsch bewerten muss, sollte sich außerdem fragen, welches Bild sie in der kurzen Zeit, die der/die Schüler/in für seine/ihre Präsentation zur Verfügung hat, bekommt (in der vorliegenden empirischen Studie war dies eineinhalb Minuten). Wie zuverlässig ist das Bild? Ist der/die Schüler/in nervös, weil das Referat aufgenommen wird? Ist die Heizung zu hoch eingestellt? Fährt gerade ein Krankenwagen mit Sirenen vorbei? Leidet der/die Schüler/in unter Bauchschmerzen? Die mündlichen Leistungen, die unter diesen besonderen Umständen erbracht werden, sind nicht unbedingt repräsentativ für seine/ihre mündliche Ausdrucksfähigkeit im Allgemeinen. Die Lehrkraft muss deshalb auf jeden Fall versuchen, die Umstände zu optimieren, damit externe Faktoren keine störende Rolle spielen können: kein Lärm, keine Hitze usw.
Eine weitere Lösung für die Verbesserung der Leistungszuverlässigkeit besteht darin, öfter die mündlichen Leistungen der Schüler/innen im Fach Deutsch zu bewerten. Dabei sollten unterschiedliche Themen behandelt werden (Wesdorp, 1981; Meuffels 1994, Van den Bergh & Meuffels, 2000). Wesdorp (1974) behauptet, dass für eine leistungszuverlässige Messung der schriftlichen Leistungen mindestens fünf Mal bewertet werden muss, Van den Bergh (1988) legt sich auf vierzehn Mal fest und Van den Berg, De Glopper und Schoonen (1988) sagen bis zu dreißig Mal. Die Bewertung sollte außerdem zur gleichen Zeit und unter den gleichen Bedingungen stattfinden. Allerdings ist dies im Unterricht praktisch unmöglich: Sechzig Abiturienten und Abiturientinnen müssen ihre mündliche Prüfung zu einem bestimmten Zeitpunkt ablegen, ihre mündlichen Prüfungen können unmöglich an ein und

demselben Tag abgelegt werden, geschweige denn in ein und derselben Stunde.

Es wird klar, dass der Begriff der „Zuverlässigkeit“ mit einigen Schwachstellen verbunden ist. Die drei unterschiedlichen Formen der Zuverlässigkeit - Stabilität, IRR und Leistungszuverlässigkeit - können die Objektivität einer Bewertung beeinflussen. Da es bei der Bewertung von Referaten bis heute keine vollständige Objektivität gibt - schließlich können die menschlichen Beurteiler (noch) nicht durch Maschinen oder Computer ersetzt werden -, muss man sich in einem solchen Fall damit begnügen, einen „vernünftigen Grad der intersubjektiven Übereinstimmung“ anzustreben (De Groot, 1961). Die Lehrkräfte müssen daher bei der Bewertung der mündlichen Prüfungen nicht in allen Fällen vollständig übereinstimmen, im Großen und Ganzen sollte dies allerdings der Fall sein.

Diese Studie beschränkt sich, u. a. aus praktischen Gründen, auf nur eine Form der Zuverlässigkeit: die IRR. Um ein angemessenes IRR-Niveau zu erreichen, ist eine Garantie erforderlich, ein Bezugsrahmen, an dem sich die Beurteiler orientieren können (De Groot, 1961). Der Bezugsrahmen in dieser Studie ist eine Skala mit einem Anker, die den Beurteilern zur Verfügung gestellt wird. Es wird erwartet, dass im Vergleich zu einer globalen Bewertung die Verwendung dieser Skala zu einer höheren IRR führen wird.

3.2 Wann ist die Übereinstimmung hoch genug?

Der Grad der intersubjektiven Übereinstimmung (IRR) lässt sich numerisch in Form einer Maßzahl zwischen 0 und 1 berechnen: Je höher die Maßzahl ist, desto zuverlässiger ist das Urteil; wenn die IRR die Maßzahl 1 hat, bedeutet dies, dass die Bewertungen der verschiedenen Beurteiler vollkommen übereinstimmen. Ist die Maßzahl 0, sind die Bewertungen total zufällig (Van der Ark, 2019).
Wie hoch dann die Maßzahl für die Zuverlässigkeit sein *muss*, ist unter anderem abhängig von dem Fach, in dem die Studie durchgeführt wird. Es versteht sich von selbst, dass die IRR höher sein sollte, wenn beispielsweise

zwei Ärzte über die Behandlung eines Schwerkranken urteilen müssen als bei einer beliebigen schriftlichen Schulprüfung. Je größer die Konsequenzen des Urteils für den Betroffenen, d. h. je schwerwiegender das Urteil, desto zuverlässiger muss dieses Urteil sein. Deswegen ist es bei wichtigen Entscheidungen - zum Beispiel bei Versetzungen oder bei der Frage, ob eine Prüfung bestanden ist oder nicht - von großer Bedeutung, dass eine IRR überhaupt berechnet werden kann. In der Schule werden oft Zensuren - und damit Bewertungen - häufig nur von einem Beurteiler vergeben. Das hat zur Folge, dass die IRR nicht ermittelt werden kann, weil dafür mehrere Beurteiler benötigt werden (Van der Ark, 2019). Es wird daher empfohlen, bei schwerwiegenden Urteilen mehrere Beurteiler hinzuzuziehen.
Da die Bewertung der Zuverlässigkeit von der Art der Beurteilung abhängt, kann keine allgemein gültige Aussage über die Höhe der IRR getroffen werden.

In der vorliegenden empirischen Studie wird die IRR mit dem (in der Testtheorie sehr bekannten) Cronbach’s Alpha berechnet. Für diese Zuverlässigkeitsmaßzahl wurden klare, weithin akzeptierte normative Kriterien entwickelt: Eine Maßzahl unter 0,5 bedeutet eine unakzeptable Zuverlässigkeit, zwischen 0,5 und 0,6 ist sie schlecht, zwischen 0,6 und 0,7 fragwürdig, zwischen 0,7 und 0,8 akzeptabel, zwischen 0,8 und 0,9 gut und über 0,9 ausgezeichnet (George & Mallery, 2003, S.231). Dieses Maß wird in dieser Studie für die Interpretation der Untersuchungsergebnisse verwendet.

Für die vorliegende Studie wurde darüber hinaus kein weiterer Versuch unternommen, eine gewisse Mindestmaßzahl für die Zuverlässigkeit zu erhalten. Die IRR wird in der ersten (globalen) Bewertungsrunde berechnet und - aufgrund des aus methodischer Sicht erforderlichen Kontrasts - anschließend nochmals in der zweiten Bewertungsrunde (anhand einer Skalenbewertung). Dabei wird davon ausgegangen, dass sich die IRR in der zweiten Runde verbessern wird.

3.3 Gültigkeit

Da diese Studie darauf abzielt, die IRR mit der Skalenbewertung zu verbessern, wird die Gültigkeit der Bewertungen nicht weiter untersucht. Dennoch ist es wichtig, den Begriff „Gültigkeit“ kurz und bündig zu behandeln, um zu verdeutlichen, dass eine Bewertung zuverlässig sein kann ohne gültig zu sein. Eine gültige Bewertung ist eine Bewertung, in der festgelegt ist, was mit dieser Bewertung beabsichtigt war. Zuverlässigkeit ist eine notwendige, wenn auch nicht ausreichende Voraussetzung für die Gültigkeit.
In der psychometrischen Literatur werden Zuverlässigkeit und Gültigkeit oft im gleichen Atemzug genannt: Kann eine Bewertung zuverlässig sein und doch ungültig? Wenn als Ausgangspunkt wieder das Beispiel der Lehrkräfte, die die mündliche Prüfung bewerten sollen, herangezogen wird, dann ist es durchaus möglich, dass sie eine zuverlässige Bewertung abgeben, weil sie weitgehend übereinstimmen. Allerdings bedeutet das nicht, dass die Bewertung gültig ist: Die Lehrkräfte können bewusst oder unbewusst von den in Kapitel 2.2 beschriebenen Störfaktoren beeinflusst worden sein. Oft bekommen Schüler/innen, die zuletzt an der Reihe sind, eine schlechtere Note als die Schüler/innen, die zuerst an der Reihe waren (Meuffels, 1994); eine solche Bewertung ist daher ungültig. Ein anderes Beispiel sind Lehrkräfte, die sich durch ihren eigenen Eindruck, den sie von einem/r guten Schüler/in haben, beeinflussen lassen. Dann kann es vorkommen, dass Lehrkräfte grammatikalische Fehler von guten Schüler/innen nicht so streng bewerten wie von schlechten Schüler/innen. Die Grammatik von guten Schüler/innen ist ausgezeichnet, aber die Botschaft ihrer Geschichte wird nicht ausreichend vermittelt. Wenn schlechte Schüler/innen dies umgekehrt machen, wer hat dann die Aufgabe besser ausgeführt und verdient daher eine bessere Note? Wenn Lehrkräfte einem/r guten Schüler/in eine bessere Note geben, verzerrt dieser Halo-Effekt die Bewertung, was dazu führt, dass die Bewertung ungültig ist. Ein weiteres Beispiel: Meuffels (1994) bezieht sich auf einen Mathematiktest, in dem nur Additionsaufgaben vorkommen; wie gültig ist ein solcher Test, wenn es darum geht, dass die Lehrkraft das mathematische Wissen der Schüler/innen messen will?

Um eine gültige Bewertung der mündlichen Leistungen im Fach Deutsch zu bekommen, darf daher nur das beurteilt werden, was beurteilt werden sollte (Meuffels, 1994; Van den Bergh & Meuffels, 2000). Das ultimative Ziel ist letztendlich, die „echten“ Sprachkenntnisse eines/r Schülers/Schülerin zu messen (Wesdorp, 1981). Soviel ist mittlerweile klar: Im Alltag ist dies oft eine äußerst komplexe Angelegenheit.

3.4 Arten der Skalenbewertung

Laut Forschungsliteratur hat die Skalenbewertung viele Varianten (siehe zum Beispiel Hsüeh Chang Chou, 1923; Fulcher, 1993), von denen jede ihre Vor- und Nachteile hat. Einige Skalenbewertungen verwenden fünf (mit zunehmender Qualität) Anker, die jeweils eine Note haben, andere besitzen nur drei oder sogar nur einen Anker. Es gibt Skalenbewertungen, bei denen die Beurteiler Noten vergeben müssen, bei anderen sind es Plus- oder Minuspunkte (besser bzw. schlechter als der Anker), ein Pluspunkt, wenn das zu bewertende Produkt etwas besser als der Standard ist, und zwei Pluspunkte, wenn es viel besser ist usw.

In Bezug auf die mündlichen Leistungen findet man in der Fachliteratur kaum Studien, die sich mit der Skalenbewertung mit Ankern befassen (es gibt zahlreiche Artikel über analytische Bewertungsmodelle, aber die Anwendung von analytischen Bewertungsmodellen scheint - entgegen den Erwartungen - empirisch für wenig Zuverlässigkeitsgewinn zu sorgen; siehe Kapitel 2.3). Wesdorp (1981) zeigt auf, wie eine Skalenbewertung in Bezug auf den mündlichen Ausdruck aussehen könnte: Auch hier muss eine Skala mit einer Reihe von Referaten (Ankern) erstellt werden, die von gut nach schlecht angeordnet sind. Bei der Bewertung versucht der Beurteiler, das Referat auf die betreffende Skala zwischen den anderen Referaten einzuordnen. Dafür gibt es zwei Varianten: (1) eine Skala mit aufgezeichneten Referaten (reale Beispiele) oder (2) eine Skala ohne reale Beispiele, sondern mit verbalen Beschreibungen der Eigenschaften eines guten, mittelmäßigen und schlechten Referats.

Ein offensichtlicher Nachteil der ersten Art der Skalenbewertung, bei der reale Referate aufgezeichnet wurden, ist, dass dies nur möglich ist, wenn die Referate, die die Skala markieren, auch tatsächlich aufgezeichnet wurden, sodass die Beurteiler die zu bewertenden Referate mit den Ankerreferaten vergleichen können. Es versteht sich von selbst, dass die Erstellung dieses Skalentyps und seine Anwendung für die Bewertung eine außerordentlich arbeitsintensive Aufgabe ist. Hinzu kommt, dass ein solcher Skalentyp nicht verwendet werden kann, wenn die Lehrkraft eine andere Präsentation aufgeben möchte: Sie müsste dann wieder eine neue Skala mit neuen Referatsankern erstellen. Die Lehrkraft kann also, wenn sie die Skala weiterhin verwenden möchte, das Thema des Referats nicht oder nur marginal ändern. Darüber hinaus ist die Wahl der Ankerreferate alles andere als einfach: Es werden sehr viele Bewertungen für sehr viele Referate benötigt, um ein zuverlässiges Urteil darüber fällen zu können, welches Referat nun gut, mäßig oder schlecht ist (Wesdorp, 1981).

Bei der Bewertung der beiden produktiven Kompetenzen der schriftlichen und mündlichen Leistungen muss der entscheidende Unterschied zwischen diesen beiden Fähigkeiten gebührend berücksichtigt werden: Im Falle schriftlicher Leistungen wurde buchstäblich etwas zu Papier gebracht und ist deshalb für die Ewigkeit erhalten. Bei mündlichen Leistungen ist das überhaupt nicht der Fall, der mündliche Ausdruck ist etwas Momentanes und Temporäres, etwas Flüchtiges. Damit ein Schülerreferat bewertet werden kann, müssen die Ankerreferate immer wieder abgespielt werden, um sicherzustellen, dass der Bezugsrahmen dem Beurteiler immer klar vor Augen steht. Das ist nicht nur arbeitsintensiv, sondern auch zeitaufwändig - ganz abgesehen davon, dass die wiederholte Präsentation ein und desselben Ankers für den Beurteiler ziemlich nervtötend sein kann. In Bezug auf den mündlichen Ausdruck liegt es deshalb auf der Hand, die Anzahl der Ankerreferate auf ein praktikables Maß zu reduzieren. In Kapitel IV wird näher erläutert, dass in dieser Studie nur ein einziges Ankerreferat verwendet wurde.

Aus den vorangegangenen Ausführungen konnte der Eindruck entstehen, dass die Skalenbewertung in der Praxis nur schwer realisierbar ist. Zugegebenermaßen gibt es offensichtliche Nachteile bei dieser Bewertungsmethode, aber die potenziellen Vorteile sollten auf keinen Fall unterschätzt werden. Der grundlegende Vorteil der Skalenbewertung - die Beseitigung einer großen Anzahl objektivitätsbedrohender Störeffekte - wurde in Kapitel 2.4 ausführlich erörtert. Kurz gesagt: Die Skalenbewertung ist konkret und gibt dem Beurteiler Halt, damit er nicht abdriftet und den Überblick verliert. Aus einer Vielzahl von empirischen Studien (Kapitel 2.3 und 2.4) zeigt sich besonders bei der Skalenbewertung von schriftlichen Leistungen ein ansehnlicher Zuverlässigkeitsgewinn (im Vergleich zu der globalen oder der analytischen Bewertung). Daraus folgen hoffnungsvolle Perspektiven für ihre Anwendung, auch bei mündlichen Leistungen.

IV Studiendesign

In diesem Kapitel werden die Methoden, Techniken und Verfahren erörtert, die in dieser empirischen Untersuchung verwendet werden. Die Entscheidungen, die in diesem Zusammenhang getroffen wurden - u. a. der Referatstyp, der Schülertyp, die Aufzeichnungsumstände, die Art der angewandten Skalenbewertung, die Frage, wie die graue Maus zustande kam, wie die Beurteiler ausgewählt wurden, welche spezifischen Anweisungen sie erhalten haben und wie sie ihre Bewertungen der Referate ausdrücken sollten - all dies wird in diesem Kapitel erläutert und - falls erforderlich - kurz begründet.
Aus methodischer Sicht muss vorab bemerkt werden, dass die aktuelle Corona-Krise (Februar-Juni 2020) großen Einfluss auf die Anweisungen der Beurteiler sowie auf die Art und Weise, wie die Bewertungen der Referate stattfanden, hatte: die Datenerfassung. Infolge dieser Krise konnten die Anweisungen und die Datenerfassung leider nicht physisch, sondern nur auf digitalem Wege stattfinden. Eine der Konsequenzen ist eine geringere Kontrolle des Bewertungsprozesses als dies aus rein methodischer Sicht wünschenswert gewesen wäre. Dies wird bei der Auswertung und Diskussion der Ergebnisse berücksichtigt. Der Mangel an strenger Kontrolle führte auch dazu, dass die vorliegende empirische Studie, wie bereits in Kapitel 2.5 angegeben, keinen prüfenden, sondern einen explorativen Charakter besitzt.

4.1 Das Stimulusmaterial: Referate

Kwakernaak (2013) argumentiert, dass der mündliche Ausdruck ein alter Begriff ist, der sich sowohl auf das Referat als auch auf ein Gespräch beziehen kann; ab 1986 wurde der unverwechselbare Begriff „Konversationskompetenz“ eingeführt. Der alte Begriff „mündlicher Ausdruck“ wurde mit zwei Formen von Übungen oder Tests assoziiert: der „Vortrag“ (in diesem Fall ein Referat, eine Präsentation, der Monolog also) und der „Buch-Test“ (in diesem Fall eine mündliche Prüfung über Literaturgeschichte oder eine Buchliste). Heutzutage gilt der Begriff „mündlicher Ausdruck“ nur für das Referat.

In dieser Studie wurde eine Mischung aus beiden Möglichkeiten gewählt: sowohl der Vortrag als auch der Buch-Test. Die Schüler/innen bekamen die Aufgabe, einen ein- bis eineinhalb minütigen Vortrag über ihr Lieblingsbuch auf Deutsch zu halten - das Buch musste nicht auf Deutsch geschrieben sein - in dem sie kurz den Inhalt des Buches wiedergeben mussten und die Frage beantworten sollten, wie sie dieses Buch verfilmen würden. Die Wahl für die Vortragsdauer von neunzig Sekunden war rein praktischer Natur. Wenn der Vortrag länger gewesen wäre, hätte die gesamte Bewertungssitzung erheblich länger gedauert und es wäre höchstwahrscheinlich viel schwieriger gewesen, Lehrkräfte für diese Studie zu gewinnen. Zudem hätten aufgrund der Dauer einer solchen Bewertung Müdigkeitseffekte bei den Beurteilern auftreten können, mit allen negativen Folgen für die Zuverlässigkeit, die das nach sich ziehen würde.

Den Schülern/Schülerinnen, deren Referate von (maximal) eineinhalb Minuten bewertet werden sollten, stand es völlig frei, ein Buch auszuwählen, über das sie kurz sprechen wollten. Die Schüler/innen konnten daher ein Buch auswählen, das sie persönlich interessierte und faszinierte. Auf diese Weise sollte die Leistungszuverlässigkeit erhöht werden. Wir wollten den Schüler/innen die Möglichkeit bieten, einen - in ihren Augen - optimalen Vortrag zu halten. Wenn sie selbst ein Buch auswählen, wissen sie aller Wahrscheinlichkeit nach selbst wirklich etwas darüber. So sollte vermieden werden, dass sie einen Vortrag über ein Buch halten mussten, das sie kaum ansprach.

Die freie Wahl lag auch im Interesse des Beurteilers. Denn wenn er 25 Mal einen Vortrag über dasselbe Buch anhören müsste, würde seine Aufmerksamkeit für inhaltliche Aspekte des Referats schnell nachlassen, was zu einem Zuverlässigkeitsverlust führen würde. Ein weiterer Nachteil, der hätte auftreten können, war das Auftreten eines Halo- und eines Sequenzeffekts. Der Halo-Effekt hätte bei einer erzwungenen Wahl auftreten können, da die Urteile hauptsächlich auf dem Präsentationsaspekt des Referats und nicht auf dem Inhalt hätten beruhen können. Auch der Sequenzeffekt hätte eine Rolle spielen können: Wenn alle Referate das gleiche Buch zum Thema gehabt hätten, wäre nicht ausgeschlossen gewesen, dass bei den Beurteilern die Neigung bestanden hätte, einen

(inhaltlichen) Vergleich mit dem vorangegangenen Referat zu ziehen (Rijlaarsdam & Bronkhorst, 1983, S.27).

4.2 Die Schülertypen

Es wurden Referate von 25 Gymnasiasten und Gymnasiastinnen der 10. Klasse bewertet, darunter dreizehn Jungen und zwölf Mädchen. Alle waren in der Altersgruppe 14 bis 16 Jahre. Alle Schüler/innen haben seit der 8. Klasse Deutschunterricht und hatten in der 8. und 9. Klasse zwei Deutschstunden von 70 Minuten pro Woche. In der Klasse sitzen 25 Schüler/innen, darunter keine sitzengebliebenen Kinder, und alle Kinder sprechen Niederländisch als Muttersprache.
Bei dieser Klasse handelt es sich um die Klasse des Verfassers dieser Masterarbeit. Nach Ansicht des Verfassers stellt diese Klasse in Bezug auf die mündlichen Leistungen im Fach Deutsch eine durchschnittliche, „normale“ Schulklasse dar, die insgesamt weder als besonders gut oder schlecht bezeichnet werden kann.

4.3 Aufzeichnungsbedingungen für das Stimulusmaterial

Alle Schüler/innen bekamen zwei Wochen Zeit für die Vorbereitung ihres Referats. Sie durften während der Aufzeichnung keine Hilfsmittel, z. B. Stichwörter, benutzen. Alle Schüler/innen haben am selben Tag und in derselben Unterrichtsstunde ihr Referat gehalten, um die Leistungszuverlässigkeit zu gewährleisten. Sie sind alle einzeln in die Klasse hereingekommen und saßen der Lehrkraft gegenüber - die Lehrkraft war der einzige Zuhörer - und die anderen Schüler/innen saßen im Flur. Die Referate wurden mit dem Telefon aufgezeichnet (die Referate wurden anschließend bearbeitet, um die Hintergrundgeräusche herauszufiltern). Die Schüler/innen hatten die Möglichkeit, die Aufnahme selbst zu starten, wenn sie bereit waren. Dann hielt der/die Schüler/in sein/ihr Referat, während die Lehrkraft nur zuhörte: Sie machte sich keine Notizen, kommentierte nichts und verhielt sich ansonsten völlig passiv. Diese Maßnahmen wurden ergriffen, damit der/die Schüler/in nicht nervös wurde und eine optimale Leistung erbringen konnte.

Nach dem Referat verließ er/sie die Klasse und der/die nächste Schüler/in konnte die Klasse betreten.

Insgesamt kann festgestellt werden, dass die Aufzeichnungsbedingungen bei der Aufnahme der Referate, dem Stimulusmaterial, so beschaffen waren, dass die Bezeichnung „normal“ zutreffend ist. Gleiches trifft für die Schulgruppe zu, deren Referate bewertet werden sollten.

4.4 Die Auswahl der Bewertungsskala in der zweiten Runde

Die Beurteiler in dieser Studie mussten die Dialoge in zwei Runden bewerten: zuerst global / ganzheitlich (nach der üblichen Schulnotenskala) und anschließend einige Zeit später mithilfe einer (Variante der) Skalenbewertung. In diesem Kapitel wird die Auswahl der Bewertungsskala in der zweiten Runde erörtert.
In der zweiten Bewertungsrunde dieser empirischen Studie sollte eine Skala verwendet werden, die aus nur einem Anker bestand; ein einziger Referenzpunkt, sodass die Bewertungsmethode übersichtlich blieb und nicht zu zeitaufwändig und umständlich wurde. Im Gegensatz zu einer Bewertung der schriftlichen Leistungen anhand einer Skalenbewertung, bei der oft fünf Ankertexte verwendet werden und diese Ankertexte bei der Bewertung jedes einzelnen Textes verwendet werden *müssen*, ist es bei der Bewertung des mündlichen Ausdrucks unmöglich, so viele Ankerpunkte zu verwenden; jedes Mal fünf Referate anzubieten ist in praktischer Hinsicht viel zu aufwendig.

Wir wollten außerdem, dass die Beurteiler ihre qualitative Bewertung eines Referats mit einem numerischen Wert ausdrückten, damit festgestellt werden konnte, inwieweit die numerischen Werte der verschiedenen Beurteiler zu ein und demselben Referat übereinstimmen. Aus dieser Perspektive fiel die Wahl auf eine Art von Skalenbewertung, die unter dem Begriff *„Stevens‘ Potenzfunktion“* (Stevens‘ power law) bekannt ist (Stevens, 1975).
Die *„Stevens‘ Potenzfunktion“* ist ein Bewertungsverfahren aus der Psychophysik, um Empfindungen, die Sinneswahrnehmungen, messen zu können. Stevens verwendete diese Methode beispielsweise für die Messung

von Geräuschen (einem Knall): Wie nehmen menschliche Beurteiler physische Reize wie den Schall wahr? Gibt es eine Eins-zu-Eins-Beziehung? Empfinden menschliche Beurteiler ein Geräusch auch doppelt so laut, wenn dieses Geräusch im physischen Sinne (in Dezibel) doppelt so laut ist?
In Stevens Experimenten bekamen Beurteiler Stimuli (d. h. Geräusche) zu hören. Ihre Aufgabe bestand darin, den Stimuli einen numerischen Wert zuzuweisen, basierend darauf, wie intensiv sie das Geräusch gehört hatten. Je lauter das Geräusch war, desto höher sollte der numerische Wert sein. Für die numerischen Werte gab es keine Ober- oder Untergrenze: Der Beurteiler selbst konnte den Wert frei festlegen.
Dem ersten Stimulus wurde ein bestimmter, fester numerischer Wert gegeben (dies konnte entweder durch den Prüfer oder den Beurteiler geschehen; schließlich wurde festgestellt, dass zuverlässigere Ergebnisse entstanden, wenn der Beurteiler, und nicht der Prüfer, dem ersten Stimulus einen Wert zuordnete). Anschließend bekamen alle anderen Stimuli, basierend auf diesem ersten Stimulus, einen numerischen Wert. Es ging also um die Wahrnehmung der Beurteiler, wie intensiv sie das Geräusch empfunden haben als Reaktion auf den ersten Stimulus.
Wir wollten in dieser empirischen Studie dasselbe Verfahren anwenden. In welchem Maße schätzen Beurteiler ein bestimmtes Referat besser, schlechter (oder gleichwertig) ein im Vergleich zu einem Ankerreferat? Die entscheidende Frage war, ob ein solches Verfahren zu einer höheren IRR führt als die ganzheitliche Bewertung der ersten Bewertungsrunde.
Der nachteilige Effekt, der bei der Verwendung von lediglich einem Anker mit ziemlicher Sicherheit vorhergesagt werden konnte, war das Auftreten des persönlichen Vergleichs: Der Beurteiler konnte, da er keine Unter- und / oder Obergrenze hatte, abdriften und extreme Werte vergeben. Bei den Ergebnissen und der Diskussion wird geprüft, ob dieser Effekt tatsächlich aufgetreten ist.

4.5 Die Auswahl der grauen Maus

Die graue Maus: Der besondere Ankerpunkt, auf den sich jeder Beurteiler in seinem Urteil beziehen sollte, ist folgendermaßen zustande gekommen: Aus

den 25 Referaten hat der Verfasser eine Auswahl aus den in seinen Augen fünf schlechtesten Referaten getroffen. Er hat dabei besonders auf drei Aspekte geachtet: (1) konnte der/die Schüler/in deutlich machen, was er/sie sagen wollte, (2) wurde die Aufgabe erfüllt und (3) konnte die Botschaft fließend ausgedrückt werden. Anschließend präsentierte er die seiner Meinung nach fünf schlechtesten Referaten einer Jury aus vier Lehrkräften, die nicht an der (eigentlichen) empirischen Studie teilnahmen. Die vier Jurymitglieder hatten alle einen Abschluss für den Lehramtstyp 4 sowie 3 bis 36 Jahre Unterrichtserfahrung. Unabhängig voneinander ordneten sie den 5 Referaten eine Rangfolge zu.

Die Aufgabe der vierköpfigen Jury war die Einordnung der Referate von schlecht bis gut. Das Ergebnis war folgende Rangfolge (das B steht für Beurteiler und das S für Schüler/in):

Tabelle 4.1: Rangfolge der fünf Referate, bewertet von vier Beurteilern

B1	S1	S2	S3	S4	S5
B2	S1	S2	S3	S5	S4
B3	S1	S5	S3	S2	S4
B4	S1	S4	S3	S4	S2

Schüler/in 3 kam einstimmig auf den 3. Platz, die IRR ist daher perfekt: Schüler/in S3 ist in der Stellungnahme der Expertenjury einstimmig der „Durchschnitt“ der fünf schlechtesten Referate und diese/r „durchschnittlich schlechte/r Schüler/in“ wird im weiteren Verlauf der Studie als „die graue Maus“ definiert. Die Voraussage war, dass das im Durchschnitt schlechteste Referat auch im Durchschnitt mit einer 5 (der Durchschnitt von dem niederländischen Notensystem) bewertet werden würde. Der Notendurchschnitt der grauen Maus (auf der Grundlage der Noten, die die Beurteiler in der ersten Bewertungsrunde für diese Leistung berechnet haben) ist eine 5,2. Das ist sehr durchschnittlich.

Diese graue Maus bekam anschließend die Maßzahl 100 zugewiesen. Stevens (1975) verwendete in seinen Studien auch die Maßzahl 100, das haben wir aufgrund der konsistenten Ergebnisse bei Stevens mit dieser Maßzahl so übernommen. Die Beurteiler haben also keine eigene Maßzahl erstellt, u. a. aufgrund des hohen Zeitaufwands und den damit verbundenen Umständen, die eine solche Methode bei zwölf Beurteilern mit sich bringen würde.

4.6 Die Beurteiler

Ursprünglich war geplant, zehn Lehrkräfte als Beurteiler für die Referate einzusetzen. Es ist unmöglich, eine generelle Aussage über die Frage zu treffen, wie viele Beurteiler in einem bestimmten Fall benötigt werden: Es hängt davon ab, was gemessen werden soll und inwieweit die Beurteiler in ihrer Bewertung übereinstimmen. Stevens (1975) behauptet, dass für eine *„einfache sensorische Skalierung“* (S.30) im Allgemeinen Gruppen aus zehn Beurteilern stabile psychometrische Ergebnisse erzielen. Tatsächlich kann erst nach der Datenanalyse eine korrekte Aussage über die Mindestanzahl an Beurteilern (bzw. Items, Messinstrumenten), die erforderlich ist, um eine bestimmte (höhere oder niedrigere) Zuverlässigkeit zu erzielen, getroffen werden.
Bei der Auswahl der Beurteiler war - aufgrund von Verallgemeinerungsmöglichkeiten - das Ziel, eine breit gefächerte Gruppe in Bezug auf Geschlecht, Alter, Unterrichtserfahrung und Lehramtsbefugnis (4. oder 3. Lehramtstyp) zusammenzustellen. Zu diesem Zweck wurden per E-Mail, Telefon und Chat insgesamt 37 Lehrkräfte angesprochen, darunter 21 Frauen und 16 Männer. Alle angesprochenen Beurteiler waren dem Verfasser bekannt. In den Gesprächen wurde den Beurteilern mitgeteilt, dass sie die Referate von Zehntklässlern zweimal beurteilen sollten; das würde insgesamt ungefähr zwei Stunden in Anspruch nehmen in einer Zeitspanne von ungefähr zwei Wochen. Schließlich sagten 14 Beurteiler ihre Teilnahme an der Studie zu, 13 Männer und 1 Frau. Sämtliche Lehrkräfte kannten die Schüler/innen, deren Referate bewertet werden sollten, nicht, sodass davon ausgegangen

werden kann, dass eine Kontamination im engeren Sinne sowie der Halo-Effekt (weitgehend) ausgeschaltet waren.
Zwölf Lehrkräfte (zwei Beurteiler haben den Verfasser nicht mehr kontaktiert) nahmen an der ersten Bewertungsrunde teil. In der zweiten Bewertungsrunde stieg die Lehrkraft B9 aus der Studie aus (dieser Beurteiler fand das gesamte Verfahren zu umständlich und meinte, dass die Schüler/innen kaum Deutsch sprechen könnten, sodass die Bewertung eines Referats sinnlos sei). Die Bewertungen der betreffenden Lehrkraft wurden in die erste Bewertungsrunde miteinbezogen; in der zweiten Bewertungsrunde wurden die fehlenden Daten als *missing value* verarbeitet.

Die folgende Tabelle enthält weitere Informationen zu den Beurteilern. B steht für Beurteiler.

Tabelle 4.2: Geschlecht, Alter, Lehramtsbefugnis (3.oder 4. Lehramtstyp) und die Jahre der Unterrichtserfahrung, pro Beurteiler

	B 1	B 2	B 3	B 4	B 5	B 6	B 7	B 8	B 9	B 10	B 11	B 12
Geschlecht	m	m	m	m	m	f	m	m	m	m	m	m
Alter	60	62	49	29	26	63	72	43	72	49	26	57
Lehramtsbefugnis	4	4	3	4	4	3	4	3	4	4	4	4
Unterrichtserfahrung	17	39	23	10	4	15	49	20	28	26	2	34

Aus Tabelle 4.2 lässt sich ableiten, dass insbesondere die etwas älteren Lehrkräfte an der Studie teilnahmen (das Durchschnittsalter einschließlich B9 beträgt 50,7 und ohne B9 48,7 Jahre). Über den Grund kann nur spekuliert werden. Möglicherweise hatten sie aufgrund ihrer größeren Erfahrung (die durchschnittliche Unterrichtserfahrung einschließlich B9 beträgt 22,9 und ohne B9 21,7 Jahre) mehr Zeit für außerschulische Aktivitäten als ihre jüngeren Kollegen.

An der Studie nahm nur eine weibliche Lehrkraft teil, die Deutschlehrerin ist in dieser Stichprobe daher deutlich unterrepräsentiert. Außerdem sind die Lehrkräfte mit einer Lehramtsbefugnis für den Typ 3 in der Minderheit. Trotz unseres Bestrebens, eine möglichst repräsentative Gruppe von Beurteilern zusammenzustellen, geben wir uns mit dieser Verzerrung zufrieden: Schließlich vertrauen wir darauf, dass die Skalenbewertung sich in allen möglichen Situationen bewähren kann und daher nicht an persönliche Hintergrundmerkmale gebunden ist.

4.7 Anleitung für die Beurteiler

Sämtliche Beurteiler bekamen die Anweisung, die Referate zweimal zu bewerten, das erste Mal global (ganzheitlich) und das zweite Mal mithilfe der „*Stevens' Potenzfunktion"*. Die Beurteiler haben sowohl in der ersten als auch in der zweiten Runde alle Referate in der gleichen Reihenfolge bewertet, um die Datenverarbeitung nicht zu kompliziert und zu zeitaufwändig zu gestalten.

4.7.1 Anleitung für die ganzheitlichen Bewertungen

Die Anleitungen für die erste Bewertungsrunde bekamen alle Beurteiler per E-Mail. Die E-Mail enthielt einen Link zu dem digitalen Formular (Google Forms), mit dem sie die Referate mit einer herkömmlichen Schulnote bewerten konnten. Die Beurteiler füllten das Formular individuell zu Hause aus; sie hatten keine Ahnung von den Bewertungen der anderen Beurteiler und konnten daher auch nicht von den anderen Beurteilern beeinflusst werden; der sogenannte Kontaminationseffekt in engerem Sinne war dadurch völlig ausgeschlossen.
Die schriftliche Anleitung lautete wie folgt:

In Teil 1 werden Sie Audio-Fragmente aus einem Referat von Zehntklässlern eines Gymnasiums hören.
Sie können es nach Ihren eigenen Kriterien beurteilen. Sie müssen lediglich eine Note gemäß der traditionellen Notengebung 1-10[1] angeben; Sie können

[1] In den Niederlanden wird von 1-10 benotet, 10=sehr gut, 1=sehr schlecht

beispielsweise eine 8 (oder höher) oder auch eine 3,1 (oder niedriger) geben. Sie können auch eine Note mit einer Zahl hinter dem Komma vergeben. Die Note muss nicht begründet werden.

Die Schüler/innen bekamen die Aufgabe, einen kurzen Vortrag über ihr Lieblingsbuch zu halten und zu beschreiben, wie sie es verfilmen würden. Das Buch musste nicht auf Deutsch geschrieben sein. Mehr Infos:

- *Die Schüler/innen haben seit 2 Jahren Deutschunterricht.*
- *Sie haben 2 Unterrichtsstunden von 70 Minuten pro Woche in der 8. und 9. Klasse gehabt.*
- *Der Vortrag sollte zwischen 60 und 90 Sekunden lang sein.*
- *Die Vorbereitungszeit war 2 Wochen*
- *Alle Schüler/innen haben den Vortrag individuell im Klassenzimmer gehalten*
- *Alle Schüler/innen haben den Vortrag in der gleichen Unterrichtsstunde gehalten.*
- *Die Schüler/innen durften keine Hilfsmittel verwenden (Stichwörter)*
- *Nur ein einziger Zuhörer war anwesend (Lehrkraft)*
- *Der Zuhörer war passiv (hat keine Notizen gemacht).*

Die Beurteiler haben sich die Referate angehört und jedes einzelne direkt im Anschluss daran global und ganzheitlich bewertet. Das Formular musste auf einmal ausgefüllt werden, es durften keine Pausen eingelegt werden. Diese strenge Prüfungsbedingungen wurden auferlegt, um den Einfluss der Störeffekte (siehe Kapitel 2.2) auf die Bewertung weitgehend zu reduzieren. Sobald die Beurteiler das Formular ausgefüllt hatten, bekam der Verfasser eine automatische E-Mail. Der Verfasser hat gewartet, bis alle Beurteiler die erste Bewertungsrunde abgeschlossen hatten (nach etwa zwei Wochen), und hat anschließend über einen Zeitraum von drei Tagen mit allen Beurteilern ein Telefongespräch geführt, um die zweite Anleitung auszuführen.

4.7.2 Anleitung für die Skalenbewertungen

Während der Anweisungen für die zweite Bewertungsrunde erhielten die Beurteiler per E-Mail einen weiteren Link zu einem digitalen Formular. Vor der eigentlichen Bewertungssitzung führte der Verfasser mit jedem Beurteiler separat ein halbstündiges Telefongespräch. Darin erläuterte er kurz, wie dieses Mal die Bewertung durchgeführt werden sollte (siehe auch die schriftliche Anleitung weiter unten).
Zusammen mit dem Beurteiler wurde drei Mal die graue Maus abgespielt: beim ersten Mal ohne Aufgabe, beim zweite Mal musste der Beurteiler notieren, was ihm an dem Referat der grauen Maus gefiel und was möglicherweise noch schlimmer hätte sein können. Beim dritten Mal notierte sich der Beurteiler die positiven Aspekte und was an dem Referat wirklich verbessert werden sollte. Nach dem zweiten und dritten Hören diskutierten der Verfasser und der Beurteiler die aufgezeichneten Punkte. Dieses ausführliche Prozedere wurde durchgeführt, damit der Beurteiler die graue Maus (bzw. die Referenz) klar und deutlich vor Augen hatte.
Anschließend begleitete der Verfasser den Beurteiler bei der Bewertung der ersten drei Referate (für alle Beurteiler die gleichen Referate). Der Beurteiler musste mithilfe der *„Stevens Potenzfunktion“* (d.h. Steven’s power law*) einschätzen*, ob das Referat besser, schlechter oder vergleichbar mit der grauen Maus war und wie viele Male schlechter oder besser. Nach jedem Referat (von den ersten drei) führten der Beurteiler und der Verfasser ein kurzes Gespräch. Sie sprachen darüber, wie der Beurteiler zu seiner Bewertung gelangt war und was der Grund dafür war. Mithilfe dieses Gesprächs wollten wir dem Bedeutungseffekt entgegenwirken. Obwohl zu erwarten war, dass jeder Beurteiler wahrscheinlich ein anderes Bild von der grauen Maus bekommen würde, (trotz *derselben* grauen Maus konnte der Beurteiler immer noch bestimmte Eigenschaften wie Grammatik oder Aussprache wichtiger als andere Eigenschaften finden) war das Ziel des Gesprächs, jeden Beurteiler zu denselben positiven und negativen Eigenschaften der grauen Maus zu lenken. Somit würde die graue Maus für jeden Beurteiler ungefähr den gleichen „Wert“ besitzen.

Nachdem der Beurteiler und der Verfasser die drei Referate zusammen bewertet hatten, konnte der Beurteiler selbstständig weiterarbeiten.
Bei der unabhängigen Bewertung sollte der Beurteiler nach jeweils drei Referaten nochmals die graue Maus anhören, damit er die Referenz wieder klar vor Augen hatte. Auf diese Weise sollte der Sequenzeffekt weitgehend eliminiert werden. Auch der Normverschiebung konnte mithilfe der grauen Maus entgegengewirkt werden: Die Beurteiler konnten ihre Bewertungen nicht mehr an das Niveau aller Referate anpassen, sondern mussten ihre Bewertungen immer wieder erneut anpassen und auf die graue Maus beziehen.

Die schriftliche Anleitung für die zweite Bewertungsrunde lautet wie folgt:

In dieser Runde bewerten Sie die Referate erneut. Diesmal mithilfe eines Ankers / Referenzpunktes: der grauen Maus. Die graue Maus hat die Maßzahl 100.
Die anderen Referate werden in Bezug auf die graue Maus bewertet. Beispiel: Wenn Sie ein Referat doppelt so gut wie die graue Maus finden, dann füllen Sie die Maßzahl 200 aus, wenn Sie es nur ein wenig besser finden, können Sie ihm beispielsweise eine 133 geben. Es gibt keine Unter- oder Obergrenze, ABER es dürfen keine negativen Bewertungen abgegeben werden. Wenn Sie ein Referat 1000 Mal schlechter als die graue Maus finden, dann geben Sie eine 0,1. Die Referate dürfen nur mit der grauen Maus und nicht mit anderen Referaten verglichen werden! Denken Sie immer daran.
Wenn Sie mit der Bewertung beginnen, ist es wichtig, die graue Maus 3 Mal anzuhören, damit sie sich als Referenzpunkt gut einprägen kann. Danach bewerten Sie 3 Referate. Anschließend müssen Sie erneut die graue Maus anhören, damit der Referenzpunkt wieder gut verankert ist. Wiederholen Sie das nach jeweils 3 Referaten.
Wenn Sie einmal eine Maßzahl angegeben haben, sollten Sie sie nicht mehr verbessern!
Möchten Sie zwischendurch eine Pause machen? Das ist möglich, aber dann sollten Sie unten auf der Liste auf „Senden“ klicken. Ihre Antworten werden dann gespeichert und Sie erhalten eine E-Mail, mit der Sie später die Liste

wieder weiter vervollständigen können. Wenn Sie nicht auf „Senden“ klicken, sind alle Antworten verloren. Wenn Sie dies zu riskant finden, können Sie auch zunächst Ihre Antworten auf Papier notieren und dann digital ausfüllen.
Haben Sie eine Pause gemacht und möchten die Bewertung fortsetzen? Dann sollten Sie zunächst wieder 3 Mal die graue Maus anhören und danach 3 Referate.
Wenn Sie fertig sind, sollten Sie die Liste nochmals durchgehen, damit sichergestellt ist, dass keine Maßzahlen fehlen.
Wenn Sie die Liste ohne Unterbrechungen ausfüllen, dauert dies ungefähr 60 Minuten.

Bei dieser Bewertungsrunde bekamen die Beurteiler die Gelegenheit, zwischendurch eine Pause einzulegen. Das geschah bewusst, denn 60 Minuten ununterbrochen konzentriert und seriös zu beurteilen erschien uns zu lang und wir wollten Ermüdungserscheinungen bei den Beurteilern vermeiden. Außerdem sollte verhindert werden, dass das ständige Abhören der grauen Maus die Beurteiler möglicherweise irritieren könnte (siehe Kapitel 2.5). Die Voraussetzung war allerdings, dass der Beurteiler nach der Pause die graue Maus wieder 3 Mal „abhören“ musste. Ob die Beurteiler diese Bedingung tatsächlich einhielten, konnte aufgrund des digitalen Kontakts nicht überprüft werden.

Die Beurteiler hatten die zweite Bewertungsrunde in einer Woche abgeschlossen, insgesamt hat die Datenerfassung drei Wochen gedauert.

Kapitel V Ergebnisse

Alle Bewertungen der Beurteiler, sowohl die der ersten Runde (ganzheitliche / globale Bewertung) als auch die der zweiten Runde (Skalenbewertung) wurden über Excel in SPSS eingegeben und anschließend analysiert. Dabei wurden die Pluspunkte, die Minuspunkte und halbe Noten der ganzheitlichen Schulnoten kodiert als 0.2, 0.8 und 0.5. Konkret: Eine 6+ wurde als 6.2, eine 6- als 5.8 und eine 6 1/2 als 6.5 kodiert.

Tabelle 5.1 enthält einige beschreibende Statistiken für jeden Beurteiler (B1 bis B12), sowohl für die ganzheitliche Bewertung (GB) als auch für die Graue-Maus-Bewertung (GMB): die durchschnittliche Bewertung (der 24 bewerteten Referate, wobei die graue Maus in den Ergebnissen der ersten Bewertungsrunde nicht enthalten ist) in der Spalte Durchschn., die Standardabweichung der 24 Bewertungen in der Spalte SD und die kleinste Maßzahl (Min) und maximale Maßzahl (Max), die der zuständige Beurteiler den 24 Referaten gegeben hat.

Tabelle 5.1: durchschnittliche Bewertung der Referate von 24 Schüler/innen (Durchschn.) pro Beurteiler (B1-B12); Standardabweichung (SD) und Minimum (Min) und Maximum (Max) für die ganzheitliche Bewertung (GB) und für die Skalenbewertung (GMB)

	GB				*GMB*			
	Durchschn.	*SD*	*Min*	*Max*	*Durchschn.*	*SD*	*Min*	*Max*
B1	*6.88*	*.76*	*5.5*	*8.0*	*139.58*	*41.33*	*80*	*220*
B2	*6.83*	*.83*	*5.0*	*8.0*	*171.67*	*64.33*	*50*	*300*
B3	*6.99*	*.82*	*5.8*	*8.8*	*213.54*	*104.83*	*33*	*400*
B4	*6.41*	*1.07*	*2.7*	*8.5*	*191.04*	*75.25*	*50*	*350*
B5	*6.00*	*.78*	*4.6*	*7.4*	*466.46*	*361.05*	*50*	*1250*
B6	*5.87*	*1.40*	*3.0*	*8.0*	*497.50*	*244.58*	*50*	*800*
B7	*6.44*	*.51*	*5.8*	*7.5*	*194.79*	*69.56*	*100*	*350*
B8	*5.00*	*.74*	*3.5*	*7*	*138.75*	*105.23*	*50*	*500*
B9	*5.66*	*.92*	*4.0*	*7.5*	*-*	*-*	*-*	*-*
B10	*5.77*	*.96*	*4.5*	*8.0*	*362.50*	*188.39*	*100*	*900*
B11	*6.78*	*.92*	*4.5*	*8.5*	*178.75*	*42.46*	*50*	*260*
B12	*6.42*	*.87*	*4.8*	*7.5*	*128.96*	*31.03*	*80*	*180*

Die Durchschnittswerte (GB) der Beurteiler zeigen große Unterschiede. Die Differenz zwischen dem niedrigsten Durchschnittswert (B8: 5.00) und dem höchsten Durchschnittswert (B3: 6.99) beträgt 1,99 Punkte, und auch bei den anderen Durchschnittswerten zeigen sich erhebliche Unterschiede. Die ursprüngliche Hypothese, nach der die Beurteiler bei ganzheitlichen Bewertungen nicht miteinander übereinstimmen, wird durch diese Daten offensichtlich bestätigt.

Wenn außerdem die kleinste und die höchste Maßzahl der GB in die Überlegungen einbezogen werden, fällt auf, dass der eine Beurteiler als geringste Maßzahl eine 2.7 vergibt, während ein anderer Beurteiler als geringste Maßzahl eine 5.8 vergibt. Auch hier also eine sehr große Diskrepanz: Was von einem Beurteiler als Ungenügend bewertet wird, bekommt von einem anderen ein Ausreichend. Wenn die Rohdaten hinzugezogen werden, stellt sich heraus, dass es sich hier um ein und dasselbe Referat handelt. Ein ähnliches - allerdings nicht so extremes - Ergebnis zeigt sich bei den Unterschieden zwischen den höchsten Maßzahlen: Auch hier kann ein signifikanter Unterschied von 1.8 festgestellt werden. Auch hier bestätigt sich offenbar unsere ursprüngliche Hypothese: Die Beurteiler bewerten anhand unterschiedlicher Standards und Werte. Sie werden bei der Benotung von vielen irrelevanten, störenden und objektivitätsbedrohenden Effekten beeinflusst (siehe Kapitel 2.2) und kommen so zu völlig unterschiedlichen Zensuren für ein und dieselbe Leistung.

Bei der Grauen-Maus-Bewertung (GMB) fällt auf, dass viele Beurteiler die Maßzahl 50 als Minimummaßzahl vergeben haben, die Übereinstimmung scheint deshalb groß zu sein. Bei der höchsten Maßzahl sind jedoch die Unterschiede signifikant. Wenn hier beispielsweise der Bereich zwischen den verschiedenen Beurteilern betrachtet wird, lässt sich feststellen, dass dieser variiert zwischen 100 und 1200. Dies ist ein großer Unterschied in Bezug auf die Streuung; es sieht so aus, als ob die Skalenbewertung auf Intervallebene nur geringe einstimmige Bewertungen ergibt.

In der folgenden Tabelle 5.2 ist sowohl für die ganzheitliche Bewertung als auch für die Skalenbewertung die IRR in Bezug auf das herkömmliche Zuverlässigkeitsmaß aufgeführt (in diesem Fall die interne Konsistenz): Cronbach’s Alpha.

Tabelle 5.2: Zuverlässigkeit der ganzheitlichen Bewertung (GB) und der Grauen-Maus-Bewertung (GMB) für die Jury aus 12 bzw. 11 Beurteilern

	Cronbach's Alpha	*Anzahl der Beurteiler*
GB	*0,93*	*12*
GMB	*0,80*	*11*

In der Tabelle 5.2 fällt auf, dass die Zuverlässigkeit (Übereinstimmung) der ganzheitlichen Bewertungen einer Jury aus 12 Beurteilern nicht nur extrem hoch ist, sondern sogar höher ist als die der Skalenbewertung einer Jury aus 11 Beurteilern (obwohl diese auch noch als „gut" bezeichnet werden kann). Dies widerspricht unseren ursprünglich formulierten Erwartungen: Es wurde schließlich behauptet, dass eine Skalenbewertung zu Zuverlässigkeitsgewinnen führen würde - Tabelle 5.2 legt jedoch das Gegenteil nahe. Damit ist unsere Hypothese faktisch widerlegt: Aufgrund theoretischer Überlegungen haben wir angenommen, dass eine Skalenbewertung die IRR tatsächlich erhöhen würde - dies war in dieser Studie allerdings nicht der Fall.

Es muss jedoch angemerkt werden, dass die Jury in der zweiten Bewertungsrunde bei der Skalenbewertung nicht aus 12, sondern aus 11 Beurteilern bestand, ein Beurteiler weniger also. Angesichts der Tatsache, dass Cronbach's Alpha nicht nur von der mittleren Korrelation zwischen den Beurteilern (je höher diese Korrelation, desto höher das Alpha), sondern auch von der Anzahl der Beurteiler abhängt (je mehr Beurteiler, desto höher das Alpha in der Regel), könnte dies die Erklärung für die geringe Zuverlässigkeit für die Skalenbewertung sein. Wenn Beurteiler (B9), der nur *missing values* für die Skalenbewertung aufweist, aus den Ergebnissen der ersten Bewertungsrunde herausgenommen werden würde, würde die Zuverlässigkeit auch 0.93 betragen. Insgesamt lässt sich sagen, dass Skalenbewertung - entgegen der Erwartungen - zu keinem Zuverlässigkeitsgewinn im Vergleich zu der „schulischen" Methode des ganzheitlichen Benotens führt.

Da die Ergebnisse völlig anders als erwartet ausfielen, ist es sinnvoll, für jeden Beurteiler die Korrelation zwischen der ganzheitlichen Bewertung (GB) und der Grauen-Maus-Bewertung (GMB) zu analysieren: Bewerten die Beurteiler einfach aufs Geratewohl (vor allem bei der Skalenbewertung GMB) oder bewerten sie in beiden Fällen in einer Art und Weise, dass sich mehr oder weniger die gleiche Leistungsreihenfolge in den 24 Referaten ergibt?

Tabelle 5.3: Korrelation zwischen der ganzheitlichen Bewertung und der Grauen-Maus-Bewertung pro Beurteiler (über 24 Referate)

	Korrelation
B1	*.862*
B2	*.684*
B3	*.763*
B4	*.753*
B5	*.640*
B6	*.844*
B7	*.779*
B8	*.643*
B9	*Missing value*
B10	*.856*
B11	*.546*
B12	*.808*

Die Korrelationen zwischen der ganzheitlichen und der „Grauen-Maus"-Bewertung pro Lehrkraft sind gemäß Tabelle 5.3 mit Ausnahme einiger weniger recht hoch: Die Korrelationen von B2, B5 und B8 liegen zwischen 0,6 und 0,7 und können daher als mittelmäßig bezeichnet werden. B11 liegt sogar unter 0,6. Diese Korrelation ist schlecht. Trotz dieser vier Beurteiler zeigen die

übrigen Beurteiler so starke Korrelationen, dass daraus die Schlussfolgerung gezogen werden kann, dass diese Beurteiler wahrscheinlich mehr oder weniger die gleichen Kriterien bei der Bewertung nach der ganzheitlichen und der Grauen-Maus-Methode hantieren. Gilt dies auch für die Beurteiler-Jury? Haben ihre Bewertungen dieselbe Bedeutung, unabhängig davon, ob sie nun ganzheitlich oder mit der Skalenbewertung benoten?

Um festzustellen, ob die (durchschnittliche) ganzheitliche Jury-Bewertung über die 24 Referate die gleiche Bedeutung hat wie die (durchschnittliche) Graue-Maus-Jury-Bewertung über die 24 Referate, wurde zunächst die Korrelation zwischen den beiden Jury-Bewertungen berechnet (siehe Tabelle 5.4).

Tabelle 5.4: Korrelation zwischen der ganzheitlichen (GB) Jury-Bewertung und der Grauen-Maus-Jury-Bewertung (GMB) über 24 Referate

	GMB
GB	*.89*

Die Korrelation in dieser Größenordnung bestätigt unsere ursprüngliche Annahme. Selbstverständlich beträgt diese Korrelation nicht genau 1 - was in streng theoretischer Hinsicht zu erwarten wäre, wenn die GB-Jury-Bewertung und die Graue-Maus-Jury-Bewertung (GMB) genau dieselbe Bedeutung hätten: Allerdings schwächt die Unzuverlässigkeit sowohl der GB-Jury-Bewertung als auch der Grauen-Maus-Jury-Bewertung (GMB) die theoretisch optimale Korrelation von 1. Korrigiert man jedoch die Korrelation zwischen der ganzheitlichen und der grauen Maus Bewertung um diesen Unzuverlässigkeitsfaktor (mit der in der psychometrischen Literatur bekannten „Korrektur für die Dämpfung“, d.h. ‚correction for attenuation‘), dann stellt sich heraus, dass die Korrelation genau gleich 1 ist.

.892 = mittlere Korrelation zwischen GB und GMB

.929 = Cronbach's Alpha der GB

.797 = Cronbach's Alpha der GMB

.892 x √ (.929 x .797) = 1.0.

Diese Schlussfolgerung ist unausweichlich: Die ganzheitlichen (Jury-) Bewertungen sind nicht von den „Graue-Maus" (Jury-) Bewertungen zu unterscheiden. Das bedeutet, dass unsere Beurteiler bei der Benotung, ob es nun um konventionelle, schulische, ganzheitliche Benotung geht oder um die Benotung nach einer bestimmten Art der Skalenbewertung, genau die gleichen Kriterien anwenden.

KAPITEL VI Auswertung

In dem vorherigen Kapitel wurde auf der Grundlage der in dieser Studie gesammelten empirischen Daten argumentiert, dass die Hypothese in dieser empirischen Studie (eine Skalenbewertung führt zu einem Zuverlässigkeitsgewinn) falsch ist. In diesem Kapitel soll das unerwartete Ergebnis (versuchsweise) erklärt werden. In diesem Zusammenhang werden einige Empfehlungen für künftige Untersuchungen zur Bewertung des mündlichen Ausdrucks ausgesprochen.

Eine der möglichen alternativen Erklärungen für das unerwartete Ergebnis dieser Studie lautet, dass unsere Probanden, die Beurteiler, „aufs Geratewohl“ bewertet haben. Allerdings ist aufgrund der hohen Korrelation zwischen der ganzheitlichen und der Graue-Maus-Bewertung (sowohl auf individueller Ebene (siehe Tabelle 5.3) als auch auf Juryebene (siehe Tabelle 5.4), diese alternative Erklärung nicht sehr plausibel und wird deshalb außer Acht gelassen. Auch der potenziell negative Einfluss der Corona-Krise auf die Art und Weise, wie die Beurteiler eingearbeitet wurden und die Art und Weise, wie anschließend die Daten gesammelt wurden, kann unmöglich eine wesentliche Rolle gespielt haben, angesichts der hohen Korrelationen zwischen den ganzheitlichen und den Graue-Maus-Bewertungen.
Was könnte die Ursache sein für das unerwartete Ergebnis dieser Studie, dass die Hypothese (die Skalenbewertung führt zu einer zuverlässigeren Bewertung als eine ganzheitliche Bewertung) sich als falsch herausgestellt hat? Laut Fachliteratur gibt es mehrere Faktoren, die möglicherweise für das unerwartete Ergebnis verantwortlich sein könnten: (1) die Maßzahl 100, die der Verfasser der grauen Maus, d. h. dem Standard, gegeben hat, (2) die Verwendung von nur einem einzigen Anker, der dazu geführt hat, dass die Beurteiler bei der Skalenbewertung fehlgeleitet wurden (also doch der Effekt des persönlichen Vergleichs) und (3) die Unterrichtserfahrung der Beurteiler. Im Folgenden werden diese Faktoren näher erläutert.

6.1 Die Referenzmaßzahl 100

Wie in den Kapiteln 3.2 und 3.3 dargestellt wurde, wies der Verfasser der grauen Maus die Maßzahl 100 zu. Diese (mehr oder weniger willkürliche) Wahl wurde nicht den Beurteilern überlassen: Mit 12 Beurteilern wäre dies eine sehr umständliche und zeitraubende Angelegenheit gewesen. Doch aus der Arbeit von Stevens (1975) folgt, dass die *„Stevens‘ Potenzfunktion“* (Stevens‘ power law) zu sehr zuverlässigen Ergebnissen geführt hätte, wenn die Beurteiler keine vorab - von dem Verfasser - vorgegebene Maßzahl verwendet hätten. Auch die Beurteiler selbst schätzten die Arbeit / das Bewerten ohne eine feste Standardmaßzahl: Sie fühlten sich unter anderem viel freier bei der Bewertung ohne Standard im Vergleich zu der Bewertung mit einer festen Maßzahl.

6.2 Persönlicher Vergleich

In Kapitel 3.2 wurde mit ziemlicher Sicherheit davon ausgegangen, dass bei der Verwendung von lediglich einem Anker der Effekt des persönlichen Vergleichs auftreten könnte: Der Beurteiler könnte, da er keine Unter- und / oder Obergrenze hat, abdriften und extreme Werte oder im Gegenteil „unverfängliche“ Maßzahlen vergeben, die nahe am Referenzpunkt liegen. Tabelle 5.1 zeigt, dass B12 nur einen Bereich von 100 nutzt, während die Bewertungen von B5 im Bereich bis zu 1200 liegen. Das Abdriften einiger Beurteiler - da eine Unter- und / oder Obergrenze nicht festgelegt wurde - spielt vermutlich bei der niedrigeren IRR in der zweiten Bewertungsrunde eine Rolle. Das spricht dafür, in künftigen Studien zur Skalenbewertung eine Unter- und Obergrenze festzulegen.

6.3 Unterrichtserfahrung

Wir denken, dass die große Unterrichtserfahrung der Beurteilergruppe (22,9 Jahre, siehe Kapitel 4.4) der Hauptfaktor für das Scheitern der Hypothese darstellt. Laut Wesdorp (1981) ist die Skalenbewertung in bestimmten, klar definierten Situationen attraktiv: nämlich in Situationen, in denen die

Beurteiler unerfahren sind oder in denen die zu bewertende Aufgabe neu ist, d. h. wenn sie zum ersten Mal in einer Lehrform eingeführt wird. „Neue“ Lehrkräfte besitzen im Vergleich zu erfahrenen Lehrkräften, die seit zwanzig Jahren Tausende Referate von Schüler/innen gehört und bewertet haben, keinen festen Bezugsrahmen. Die „alten Hasen“ haben sich im Laufe der Jahre einen klaren Bezugsrahmen geschaffen, auf den sie bei einer komplexen Bewertungsaufgabe wie der Bewertung von Referaten immer zurückgreifen können. Angesichts der hohen durchschnittlichen Unterrichtserfahrung unserer Beurteilergruppe halten wir es für durchaus plausibel, dass die IRR (0.93) der ersten ganzheitlichen Bewertungsrunde aus diesem Grund so hoch ist. In der zweiten Bewertungsrunde, der Skalenbewertung, hingegen mussten alle Beurteiler eine völlig neue, für sie unbekannte Bewertungsmethode für die Referate anwenden. Sie mussten also einen völlig neuen Referenzrahmen nutzen und konnten nicht auf ihren vertrauten Referenzrahmen zurückgreifen. Diese Tatsache ist unserer Ansicht nach der plausibelsten Erklärung für die relativ niedrigere IRR (0.80) bei der Skalenbewertung.

6.4 Vorschläge für weiterführende Untersuchungen

Zwei Aspekte könnten in weiterführenden Untersuchungen vertieft werden. Erstens ist es empfehlenswert, den Beurteilern die Möglichkeit zu bieten, der grauen Maus selbst eine Standardmaßzahl zuzuweisen und alle anderen Referate anhand dieser Maßzahl zu bewerten. Selbstverständlich würde dies, abhängig von der Anzahl der Beurteiler, zu einer größer angelegten und längeren Studie führen als die vorliegende Studie, für die nur ein begrenzter Zeitrahmen und Raum sowie eingeschränkte Mittel zur Verfügung standen. Zweitens wäre es interessant, die vorliegende Studie mit einer Beurteilergruppe mit wenig Unterrichtserfahrung (z. B. weniger als fünf Jahre) zu wiederholen und zu untersuchen, ob die Skalenbewertung in dieser ausgewählten Gruppe zu einer höheren IRR führt. Auf Wunsch kann sie mit einer Gruppe von Beurteilern mit viel Unterrichtserfahrung (z. B. mehr als zwanzig Jahre) verglichen werden.

Literaturverzeichnis

De Glopper, K. (1989). Schrijven beschrijven. Dissertation

De Groot, A. (1966). Vijven en zessen. Groningen: Wolters-Noordhoff

De Groot, A. (1961). Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen. ‘s-Gravenhage: Mouton

Elving, K. & Van den Bergh (2015). Gewicht in de schaal. *Levende talen tijdschrift*, 16(2), 26-36

Fulcher, G. (1993). The construction and validation of rating scales for oral tests in English as a foreign language. Dissertation

George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.). Boston: Allyn & Bacon.

Hormann, H. & Verbeek, J. (1977). Correctie Nederlands Eindexamen VWO. *Levende Talen*, 143-144.

Hsueh Chang Chou, H. (1923). The Measurement of Composition Ability. New York: (z.u.).

Kreeft, H., Luyten, T. & Schreuder, K. (1978). De normaalfunctionele toets. Een betrouwbare CITO-reactie op het CMM-advies over het eindexamen Nederlands in het vwo, het havo en het mavo. *Levende Talen*, 99-108.

Kwakernaak, E. (2013). Didactiek van het vreemdetalenonderwijs (eerste druk, tweede oplage). Bussum: Coutinho

Meuffels, B. (1994). De verguisde beoordelaar. Amsterdam: Thesis Publishers.

Meuffels, B. (1978). Effect-onderzoek taalvaardigheid. *Spektator*, 47-61.

Pollmann, E., Prenger, J. & De Glopper, K. (2012). Het beoordelen met behulp van een schaalmodel. *Levende talen tijdschrift*, 13(3), 15-24.

Pullen, T. (2012). Bij wijze van schrijven. Dissertation

Rijlaarsdam, G. & Blok, H. (1981). Beoordeling van schrijfprodukten door leerlingen: theorie en praktijk. *Levende Talen*, 753-766.

Rijlaarsdam G. & Bronkhorst H. (1983). Beoordelen van spreekbeurten. Amsterdam: Stichting centrum voor onderwijsonderzoek van de universiteit van Amsterdam.

Schoonen, R. (1991). De evaluatie van schrijfvaardigheidsmetingen. Dissertation

Stevens, S.S. (1975). Psychophysics Introduction to its perceptual, neural, and social prospects. New York: John Wiley & Sons, Inc.

Van den Bergh. H. (1988). Directe metingen van schrijfvaardigheid: validiteit en taakeffecten. In F. van Emeren en R. Grootendorst (Red.), *Taalbeheersing in ontwikkeling* (pp. 370-378). Dordrecht: Foris Publications

Van den Bergh, H. (1988). Examens geëxamineerd. ‘s Gravenhage: S.V.O.
Van den Bergh, H. & Meuffels, B. (2000). Schrijfvaardigheden en schrijfprocessen. In A. Braet (Red.), *Taalbeheersing als communicatiewetenschap* (pp. 122-154). Bussum: Coutinho

Van den Ende, J. (1954). Cijfers op de middelbare school. *Pedagogische Studiën*, 31, 69-86.

Van den Ende, J. (1954). Cijfers op de middelbare school. *Pedagogische Studiën*, 31, 112-129.

Van der Ark, L., & ten Hove, D. (2019). Zijn we het eens? Interbeoordelaarsbetrouwbaarheid in de pedagogiek en het onderwijs. Pedagogische Studiën, 95(5/6), 361-371.

Van Schooten, E. (1988). De constructie van van een meerkeuzetoets voor het meten van schrijfvaardigheid. Amsterdam: Stichting Centrum voor Onderwijsonderzoek van de Universiteit van Amsterdam: Stichting Kohnstamm Fonds voor onderwijsresearch.

Wesdorp, H. (1981). Evaluatietechnieken voor het onderwijs. ‘s-Gravenhage: Staatsuitgeverij.

Wesdorp, H. (1978). Het meten van productief-schriftelijke taalvaardigheid. Dissertation