

Die Konsistenz von kontinuierlichen Bewertungen und Gesamturteilen in Live-Performances

Master-Arbeit vorgelegt im Rahmen der Master-Prüfung
für den Masterstudiengang Lehramt an Gymnasien
im Teilstudiengang Musikerziehung

von
Carolin Scholle

Osnabrück, 15. Juni 2015

Erstprüfender: Prof. Dr. Christoph Louven
Zweitprüfende: Dr. Franziska Olbertz

Inhaltsverzeichnis

Abbildungsverzeichnis	5
Tabellenverzeichnis	6
1 Einleitung	7
2 Forschungsstand	10
2.1 Definitionen	10
2.2 Überblick zur Performanceforschung	12
2.3 Theorien zur Performancebewertung	15
2.4 Empirische Befunde zu Performancebewertung	21
2.4.1 Visuelle Einflüsse auf die Performancebewertung	21
2.4.2 Konsistenz von Performancebewertungen	24
2.4.3 Bewertungsskalen	27
2.5 Kontinuierliche Datenerhebung	30
2.5.1 Sample-Rate und Latenzzeiten	32
2.6 Zusammenhang zwischen kontinuierlichen und retrospektiven Wer- tungen	34
2.6.1 Peak-End-Rule	35
2.6.2 Primacy Effekt	39
2.6.3 Evolution der Bewertung	41
3 Ziele und Hypothesen	45
4 Methode	49
4.1 Rahmen des Experiments	49
4.2 Stichprobe	49
4.2.1 Musiker	49
4.2.2 Publikum	50
4.3 Erhebungsinstrumente	51
4.3.1 Fragebogen für die Musiker	51
4.3.2 Interface zur Echtzeit-Datenerhebung	52
4.3.3 Fragebogen für das Publikum	54
4.4 Ablauf	55
4.5 Beobachtungen im Experiment	56

5	Ergebnisse	58
5.1	Konsistenz der Gesamtbewertungen	58
5.1.1	Durchschnittliche Gesamtbewertungen und Streuung	58
5.1.2	Konsistenz der Bewertungen	60
5.1.3	Einfluss von Freundschaftsgrad, Präferenz, Bekanntheit und eingeschätztem Schwierigkeitsgrad	62
5.1.4	Unterschiede zwischen Studierenden und Dozenten sowie zwischen den Geschlechtern	65
5.1.5	Expertise als Einflussgröße	67
5.1.6	Vergleich zur Selbsteinschätzung der Musiker	68
5.2	Analyse der Bewertungsverläufe	71
5.2.1	Methodische Vorüberlegungen	72
5.2.2	Innere Konsistenz der kontinuierlichen Bewertungen	80
5.2.3	Konsistenz der Bewertungsverläufe und Performanceanalyse mit Hilfe der kontinuierlichen Bewertungen	82
5.3	Zusammenhang zwischen kontinuierlichen und retrospektiven Bewertungen	89
5.3.1	Methodische Vorüberlegungen	90
5.3.2	Vergleich der Modelle	93
5.3.3	Einfluss des Bühnenauftrittsverhaltens	101
6	Zusammenfassung und Ausblick	103
	Literatur	108
	Anhang	113

Abbildungsverzeichnis

2.1	Modell des Musikbewertungsprozesses	17
4.1	Zusammensetzung des bewertenden Publikums	51
4.2	emoTouch-Session	53
4.3	Webview mit Online-Fragebogen	53
5.1	Bewertungen der fünf Performances im Vergleich	59
5.2	Verteilung der Wertungen für Ensemble 5	60
5.3	Verteilung der Wertungen für Ensemble 2	61
5.4	Streudiagramm: Bewertung und Bekanntheitsgrad bei Stück 2	65
5.5	Histogramm: Bewertungsdifferenzen	69
5.6	Streudiagramm: Publikumsbewertung und Abweichung der Selbst- einschätzung von der Publikumsbewertung	70
5.7	Die Bewertungen aller Probanden zu Performance 4	73
5.8	Die z-transformierten Bewertungen aller Probanden zu Performance 4	73
5.9	Mittelwert \pm Standardabweichung aller Bewertungen zu Performance 4	76
5.10	Mittelwert \pm Standardabweichung aller z-transformierten Bewer- tungen zu Performance 4	76
5.11	Mittelwert, Median und Quartile der Bewertungen zu Performance 4	78
5.12	Mittelwert, Median und Quartile der z-transformierten Bewertungen zu Performance 4	79
5.13	Mittlere Bewertung \pm Standardabweichung von Performance fünf im Zeitverlauf	83
5.14	Mittlere z-transformierte Bewertung \pm Standardabweichung von Performance fünf im Zeitverlauf	84
5.15	Mittlere Bewertung \pm Standardabweichung von Performance zwei im Zeitverlauf	88
5.16	Mittlere Bewertung \pm Standardabweichung von Performance drei im Zeitverlauf	88
5.17	Mittlere absolute Abweichung von der Gesamtbewertung	96
5.18	Differenzen der Bewertungen zu einzelnen Zeitpunkten und der Gesamtbewertung	97

Tabellenverzeichnis

4.1	Programm am Vorspielabend	50
5.1	Vergleich der Inter-Rater-Reliabilität	62
5.2	Korrelationen zwischen der Gesamtbewertung und der Präferenz für das Stück	63
5.3	Standardabweichung der Zeitreihen	82
5.4	Definitionen der betrachteten Einflussfaktoren	93
5.5	Korrelationen der Bewertungen zu einzelnen Zeitpunkten mit der Gesamtwertung	95
5.6	Interkorrelationen der unabhängigen Variablen	100
5.7	Korrelationen bestimmter Zeitpunkte sowie der Gesamtbewertung mit dem Bühnenaufgang	102

1 Einleitung

Die Bewertung von musikalischen Performances ist beispielsweise bei Wettbewerben und an Universitäten eine häufige Aufgabe. Eine Jury entscheidet darüber, wer den Preis gewinnt oder wer überhaupt zum Musikstudium zugelassen wird. Bei allen Auftritten vor Publikum setzen sich Musiker der Bewertung der Zuhörer aus, auch wenn diese für sie dann möglicherweise weniger ausschlaggebend ist. Nach einem Konzert wird schließlich fast immer diskutiert, wie es gefallen hat – und dabei sicher nicht vergessen, beispielsweise zu loben, wie locker der Pianist die schwierigsten Stellen spielt, oder anzumerken, wie hochnäsig er am Ende bei der Verbeugung herüberkam.

Es gibt im Alltagsverständnis viele Ansichten darüber, wie solche Meinungen gebildet werden, mit einem Spektrum von der (angeblichen) Erfahrung, dass sich die Prüfer immer bis auf geringe Abweichungen einig seien, bis hin zum Vorwurf, die Noten seien willkürlich. Auch über die Entstehungsprozesse existieren verschiedene Hypothesen, wie etwa, dass sowieso nach den ersten paar Takten die Bewertung bereits feststehe. Alternativ wird vermutet, die Juroren gehen zunächst von der Bestnote aus und ziehen Punkte ab für das, was schlecht läuft, oder umgekehrt.

Einerseits wird die Frage danach, wie zuverlässig und objektiv Bewertungen von Live-Performances sind deutlich. Würde eine andere Jury den gleichen Sieger küren? Oder würden die Prüfer, wenn sie morgen bessere Laune haben, anders bewerten? Andererseits zeigen die genannten Beispiele bereits, dass die Bewertung einer musikalischen Darbietung eine Zeitkomponente hat. Wie lange hat ein Musiker Zeit, die Juroren zu überzeugen, bis sie sich ihre Meinung gebildet haben? Da die Musik selbst Zeit benötigt, um zu existieren, ist es logisch, dass auch die Bewertung dieser Musik ein zeitabhängiger Prozess sein muss. Sie entsteht nicht erst nach dem Hören einer Performance, sondern bereits während dessen und kann sich im Zeitverlauf verändern. Nur die Note, die später festgehalten wird, hat diese Zeitkomponente nicht mehr.

Obwohl diese Bewertungsprozesse alltäglich in vielen Institutionen sind, ist das Thema wenig präsent in aktueller musikpsychologischer Forschung. Der zentrale Gegenstand der Performanceforschung sind die Musiker und deren Performance in all ihren Aspekten und die Fragestellungen sind meistens daraufhin ausgerichtet,

wie Musiker ihren Auftritt optimieren können. Es werden visuelle Einflüsse auf die Bewertung von musikalischen Darbietungen untersucht, aber selten finden die Studien in Live-Situationen statt, sondern stattdessen wird den Probanden nur ein Video gezeigt. Der Bewertungsprozess an sich wird dabei kaum hinterfragt und der Blick ist insgesamt nicht auf den Rezipienten gerichtet. Speziell bleibt bis auf in wenigen Ausnahmen, die alle auf Live-Performances verzichten, die Zeitkomponente der Meinungsbildung unberücksichtigt.

Das Ziel dieser Arbeit ist es, einen Beitrag zu leisten, diese Lücke zu schließen. Es soll untersucht werden, wie zuverlässig und einheitlich Zuhörer musikalische Performances bewerten, und dabei soll weder auf die Live-Situation noch auf die Berücksichtigung der Zeitkomponente verzichtet werden. Zuverlässig bzw. einheitlich wird sowohl auf das Publikum als Gruppe bezogen, als auch auf den einzelnen Rezipienten im Hinblick darauf, ob sein abschließendes Urteil und seine Bewertungen im Zeitverlauf zusammen passen. Konkret sollen folgende drei Forschungsfragen beantwortet werden:

1. Wie konsistent sind die retrospektiven Gesamturteile?
2. Wie konsistent sind die kontinuierlichen Bewertungen?
3. Wie hängen kontinuierliche und retrospektive Wertung zusammen?

Die erste Frage bezieht sich darauf, wie einheitlich das Publikum als Gruppe bewertet. Damit sind nicht ausschließlich gleiche absolute Bewertungen sondern auch relative Übereinstimmungen gemeint. Beispielsweise könnte ein Rater eine eher positive Bewertungstendenz haben, ein anderer eine negativere, aber sie können sich dennoch einig darüber sein, welche Performance in einer Rangordnung welchen Platz erhält und wie groß die Abstände zwischen den Rängen sind. Außerdem gehört zu dieser Frage auch das Finden oder Ausschließen von Untergruppen innerhalb des Publikums, die unterschiedlich bewerten, und inwiefern, die Selbsteinschätzungen der Musiker ebenfalls konsistent mit den Publikumsbewertungen sind.

Frage zwei bezieht sich ebenfalls auf die Konsistenz, mit der das Publikum als Gruppe bewertet, betrachtet allerdings die zeitabhängigen Bewertungen. Sowohl die momentane Konsistenz zu einzelnen bestimmten Zeitpunkten als auch die zeitlich übergreifende durchschnittliche Konsistenz soll dabei untersucht werden. Eine weitere Teilfrage ist, inwiefern die Bewertungsverläufe über die Zeit interindividuell ähnlich sind und, wenn ja, wie diese Verläufe typischerweise aussehen. Die Aspekte

lassen sich zudem anhand von Videoaufzeichnungen auf Zusammenhänge mit der Performance untersuchen.

Die dritte Frage bringt die kontinuierlichen und retrospektiven Bewertungen in Verbindung und bezieht sich darauf, wie konsistent jede einzelne Person zu den verschiedenen Zeitpunkten bewertet. Es soll herausgefunden werden, wie sich die Gesamtbewertungen aus den kontinuierlichen ergeben bzw. wie groß der Zusammenhang der beiden Bewertungsformen ist.

Nicht Ziel dieser Arbeit ist es, die Performance selbst zu untersuchen und aus den Bewertungen des Publikums Tipps für die Musiker abzuleiten. Zusammenhänge zu den musikalischen Darbietungen werden lediglich hergestellt, um mögliche Erklärungen für die Reaktionen des Publikums zu finden.

Um diese Fragen beantworten zu können, wird eine empirische Studie in einer konzertähnlichen Vorspielsituation durchgeführt, bei der das Publikum die gehörten Performances bewertet. Die Bewertungen werden kontinuierlich während der Auftritte erhoben, indem die Zuhörer auf iPads auf einer Skala durchgehend markieren, wie gut oder schlecht sie die Performance gerade finden. Außerdem werden nach jedem Stück Gesamtbewertungen abgegeben. Die musikalischen Darbietungen werden dabei auf Video aufgezeichnet um später eine Zeitreferenz zu haben und Zusammenhänge herstellen zu können.

Die Arbeit gliedert sich in einen theoretischen und einen empirischen Teil. Zunächst wird im theoretischen Teil (Kapitel 2) der Forschungsstand dargestellt, wobei nach einem kurzen Überblick die Themen Performancebewertung und kontinuierliche Selbstauskünfte als Methode in der Musikpsychologie betrachtet werden. Kapitel 3 bildet den Übergang zum empirischen Teil und konkretisiert die Ziele mit aus dem theoretischen Teil abgeleiteten Hypothesen, die überprüft werden sollen. Im empirischen Teil wird erst die methodische Vorgehensweise dargestellt (Kapitel 4) und dann werden, untergliedert nach den drei Forschungsfragen, die Ergebnisse der Datenanalyse präsentiert (Kapitel 5). Zum Schluss werden die Ergebnisse zusammengefasst, in den theoretischen Kontext eingeordnet und Schlussfolgerungen für weitere Forschung diskutiert (Kapitel 6).

2 Forschungsstand

In diesem Kapitel werden zunächst wichtige Begriffe definiert und ein allgemeiner Überblick über die Gegenstände der Performanceforschung wird gegeben. Anschließend werden spezifischere Theorien und empirische Befunde zur Performancebewertung vorgestellt, allerdings noch überwiegend ohne die Zeitkomponente zu berücksichtigen. Die letzten beiden Abschnitte befassen sich mit Echtzeitverfahren und darauf aufbauend mit Theorien, wie kontinuierliche Bewertungen und retrospektive Urteile zusammenhängen könnten.

2.1 Definitionen

In diesem Abschnitt werden die beiden zentralen Begriffe *Performance* und *Konsistenz*, die bereits im Titel dieser Arbeit verwendet wurden, definiert. Was unter *kontinuierlichen* Bewertungen verstanden wird, ist vielmehr eine methodische und keine rein begriffliche Frage und wird daher in Abschnitt 2.5 erläutert.

Eine musikbezogene Performance-Definition wird wie folgt von Kopiez (2010) übernommen:

Der Begriff „Performance“ (engl. Darbietung, Leistung) bezeichnet in Bezug auf die Musikforschung alle Aspekte der „Interpretation“ (lat. Auslegung) und der besonderen Leistungen eines Musikers. (Kopiez, 2010, S. 367)

Diese Definition ist gleichzeitig nah an der wörtlichen Übersetzung, aber konkret auf Musik bezogen. Dabei ist im Rahmen dieser Arbeit nur der Leistungsaspekt wichtig, da sich hieraus ergibt, dass eine Bewertung, beispielsweise in Form einer Note, möglich ist, während der Interpretationsaspekt sich eher auf qualitative Unterschiede bezieht und hier nicht näher betrachtet wird. Dieser wird nur relevant, wenn er als Teil der Leistung mit bewertet wird. Die Bezeichnung „besondere Leistungen eines Musikers“ ist sehr offen gehalten, sodass nicht ausschließlich die klingende Musik sondern auch das gesamte Auftrittsverhalten miteinbezogen werden kann. Somit können neben der musikalischen Leistung, wie beispielsweise

dem Spielen der richtigen Töne oder angemessener Agogik, auch außermusikalische, visuelle Aspekte wie die Spielbewegungen, Blickkontakt zum Publikum u. v. m. die Bewertung beeinflussen. Eine engere, rein auf die klingende Musik bezogene Auslegung des Begriffs würde den aktuellen Themen der Performanceforschung nicht gerecht und kann von Kopiez (2010) nicht gemeint sein, da auch sein Überblick etwa Studien zu visuellen Einflüssen auf die Performancebewertung beinhaltet. In dieser Arbeit soll der Begriff so weit gefasst werden, dass der Auf- und Abgang auf bzw. von der Bühne mit als Teil der Performance betrachtet werden, was von Platz (2014) theoretisch wie empirisch als sinnvollere zeitliche Begrenzung gegenüber einer Einschränkung auf die reine Spielzeit nahegelegt wird. Somit ist also jegliches Musikerverhalten, das auf der Bühne stattfindet, Teil der bewerteten Leistung.

Der Begriff *Konsistenz* ist dem lateinischen Wort „*consistere*; sich hinstellen, hintreten, standhalten, fort dauern“ entlehnt und die Bedeutung des Wortes von *konsistent* wird mit „fest, in sich stimmig“ angegeben (Kluge, 2011).

Im Rahmen dieser Arbeit wird jedoch folgende spezifischere Definition aus der Psychologie zugrunde gelegt. In der Differentiellen Psychologie und der Persönlichkeitspsychologie wird unter der *Konsistenz des Verhaltens* verstanden, wie konstant bzw. stabil menschliches Verhalten über Zeit oder Situationen hinweg bleibt (Häcker, 2013, S. 918). Auf Performancebewertungen bezogen bezeichnet konsistentes Verhalten, wie genau eine Person die gleiche Leistung immer gleich bewertet. Demnach kann die Konsistenz des Bewertungsverhaltens konkreter als eine Wiederholbarkeit von Bewertungen untersucht werden. Dabei kann einerseits der Zeitpunkt variieren, Bewertungen also während und direkt nach der Performance sowie mit größerer zeitlicher Verzögerung vorgenommen werden. Andererseits variiert bereits in diesem Beispiel auch die Situation dadurch, dass die Bewertung entweder während der Performance, also während des Zuhörens, vorgenommen wird oder nach der Performance aus der Erinnerung heraus. Als weitere situative Veränderungen sind verschiedene Kontexte einer Performance zu sehen, beispielsweise könnten Bewertungen in einem Konzert anders ausfallen als in einer Prüfung, aber dieser Aspekt spielt im Folgenden keine Rolle. *Konsistenz des Verhaltens* kann somit synonym zu Reliabilität im Sinne von Wiederholbarkeit benutzt werden. *Reliabilität* bezeichnet die Zuverlässigkeit und Genauigkeit eines Messinstruments (Sedlmeier & Renkewitz, 2013, 71ff). Wird ein gleichbleibendes Objekt immer gleich gemessen, misst das Instrument reliabel. Übertragen auf Performancebewertungen von dem Publikum, werden die Zuhörer zu Messinstrumenten und die Performance selbst zu dem zu messenden Objekt. Bewertet also ein Zuhörer die gleiche Performance immer gleich,

bewertet er reliabel. Im Kontext von Live-Performances ist es zwar unmöglich, die gleiche Performance mehrfach zu hören, aber beispielsweise ist ein Vergleich von Bewertungen zur gleichen Performance, die zu verschiedenen Zeitpunkten vorgenommen wurden, denkbar um die Reliabilität zu messen. Praktikabel ist insofern die Sichtweise, das ein Zuhörer reliabel bewertet, wenn er eine gleich gute Performance immer gleich bewertet, sich also beispielsweise nicht davon beeinflussen lässt, ob er direkt vorher jemand besseres oder schlechteres gehört hat. Für die Konsistenz des Verhaltens werden im folgenden die Begriffe intraindividuelle Konsistenz bzw. intraindividuelle Reliabilität verwendet, um zu verdeutlichen, dass sie sich auf eine Person bezieht und nicht auf Unterschiede bzw. Übereinstimmungen zwischen verschiedenen Zuhörern.

Für die Konsistenz, mit der das Publikum insgesamt als Gruppe bewertet, wird der Begriff der *inneren Konsistenz* aus der Testtheorie zu Grunde gelegt, der dort grundsätzlich synonym mit Reliabilität verwendet wird (Wirtz, 2014, S. 918). Die *innere Konsistenz* eines Messverfahrens, beispielsweise eines psychologischen Tests, bezeichnet die Homogenität der einzelnen Testteile und wird an den Interkorrelationen der Teile bzw. den Korrelationen der Teile mit dem Gesamtergebnis gemessen (s. Mikula, 2013, S. 918). Übertragen auf die Situation einer Performancebewertung durch das Publikum entspricht hier das gesamte Publikum dem Messverfahren und jeder Zuhörer einem Testteil. Die zu messenden Objekte sind hier die Performances der Musiker und wären bei einem psychologischen Test hingegen die teilnehmenden Subjekte. Die *innere Konsistenz* ist somit ein Maß dafür, wie homogen die einzelnen Zuhörer musikalische Performances bewerten, also wie stark die Bewertungen untereinander korrelieren und wie stark jede einzelne Bewertung mit der Gesamtbewertung korreliert. Die Gesamtbewertung könnte dabei beispielsweise als Mittelwert oder Summe aller Bewertungen berechnet werden. Synonym zu *innerer Konsistenz* werden im Folgenden auch die Begriffe interindividuelle Reliabilität und Inter-Rater-Reliabilität verwendet, also die Zuverlässigkeit mit der verschiedene Subjekte vergleichbar bewerten.

2.2 Überblick zur Performanceforschung

Innerhalb der Performanceforschung steht überwiegend die musikalische Darbietung selbst im Fokus, aber, wie Platz (2014) bereits feststellte, selten die Rezipienten. Die Analyse von Performancebewertungen hat eine Randstellung und Musikwahrnehmung wird überwiegend nicht im Kontext von Live-Musik untersucht.

Windsor (2009) schreibt beispielsweise zusammenfassend über Messung und Modelle von Performance, richtet dabei aber den Blick auf die Darbietung selbst und bezieht die Wahrnehmung auf Seiten des Publikums nicht mit ein. Kopiez (2005) hingegen inkludiert in seinem Überblicksartikel über experimentelle Interpretationsforschung, die er mit der *Performance*-Forschung im Englischen gleichsetzt, auch einen Abschnitt, der sich explizit mit der Bewertung der Performance-Qualität befasst. Als entscheidender Faktor für eine gute Performancebewertung stellt er Kohärenz heraus, die darauf bezogen wird, dass ein großer musikalischer Bogen hinsichtlich der Tempo- und Dynamikentwicklung gespannt wird, der über einzelne Phrasen weit hinaus reicht. Im Zusammenhang damit wird dargestellt, dass bzgl. der angewandten Übestrategien vor einem Auftritt diejenigen erfolgreicher seien, die zumindest ab einem gewissen Zeitpunkt beinhalten, entsprechend lange Abschnitte am Stück zu spielen und die Stücke nicht mehr überwiegend in Details zu zerlegen. Dies sei wichtig, damit der große, dadurch auch gut bewertete Zusammenhang über das gesamte Stück entstehen kann. Außerdem erläutert Kopiez (2005), dass Performer ihren Auftritt oft sehr fehlerzentriert betrachten, der Großteil der Fehler vom Zuhörer aber tatsächlich nicht wahrgenommen werde.

Die wenige Forschung, die das Publikum bei Live-Konzerten im Blick hat, untersucht vielmehr Sozialstrukturen der Zuhörer, beispielsweise für Konzerte verschiedener Genres, allerdings nicht deren Erlebnis des Konzertes an sich (vgl. Neuhoff, 2007). Ebenso betrachten Kalies, Lehmann und Kopiez (2008) die Publika verschiedener Musikszene eher mit dieser Fragestellung. Allerdings stellt das Thema in diesem Artikel zu „Musikleben und Live-Musik“ nur einen Aspekt dar und darüber hinaus geht es um Amateurmusiker und Musizieren in der Freizeit, sowie um die Bewertung von Live-Musik. Einerseits werden Unterschiede zwischen journalistischen Kritiken und Bewertungen in Wettbewerben erläutert, wobei bzgl. Benotung oder Rankings in Wettbewerben in empirischen Untersuchungen nur mäßige Übereinstimmungen der Juroren gefunden wurden. Eine genauere Analyse dieser Studien findet sich in Abschnitt 2.4.2. Andererseits thematisieren Kalies et al. (2008) die Interaktion von der gehörten Musik und dem visuellen Eindruck, wobei vorwiegend der Einfluss des Visuellen auf die Performancebewertung im Zentrum steht. Es werden empirische Ergebnisse zum Einfluss von Mimik und Gestik bzw. generell Körperbewegungen zur Unterstützung der Kommunikation, Attraktivität, Kleidung u. v. m. vorgestellt. Außerdem wird herausgestellt, dass Geschlechterstereotype die Bewertung beeinflussen. Beispielsweise könnte, ob das Instrument für das Geschlecht typisch ist, einen Einfluss auf die eingeschätzte Kompetenz haben

und bei Playback-spielenden Pianisten wurde bei gleicher gehörter Aufnahme die Interpretation von Pianistinnen mit anderen Attributen beschrieben, als die von Pianisten. Zu Bewegungen von Musikern auf der Bühne hat auch Davidson (2009, S. 374) bereits viele Ergebnisse zusammengefasst, bzgl. der Rezeption beispielsweise, dass Bewegungen akkurate Wahrnehmung der musikalischen Intention zulassen und generell eine Möglichkeit zur Teilhabe an der Performance bieten. Im Rahmen der Bewertung von musikalischen Live-Performances ist der Einfluss des Visuellen dementsprechend der bereits am besten empirisch erforschte Bereich. Zumindest teilweise sind Live-Situationen untersucht worden und ansonsten Videos, da reine Audio-Aufnahmen für diese Fragestellung ohnehin nicht verwendet werden. Einige Ergebnisse werden in Abschnitt 2.4.1 noch genauer vorgestellt, da sie auch für die Fragestellung dieser Arbeit relevant sind.

Im Kontext von Musikwahrnehmung sind einerseits Musikkognition und Verarbeitung von auditiven bzw. audiovisuellen Signalen, andererseits Musik und Emotionen, wichtige Forschungsthemen, wobei hier praktisch nie Live-Musik als Stimulus eingesetzt wird. Schlemmer (2005) erklärt, dass trotz unterschiedlicher Verarbeitungszeiten und Ausbreitungsgeschwindigkeiten der beiden Signale, Audio und Visuelles in der Wahrnehmung nicht voneinander trennbar ist. Juslin und Timmers (2010) fassen Ergebnisse zu Ausdruck und Kommunikation von Emotionen in Musik zusammen, zwar nicht im Kontext von Performancebewertungen, aber es könnte durchaus ein Beurteilungskriterium sein insofern, dass von einer guten Performance erwartet wird, Emotionen zu übermitteln. Darüber hinaus wird die Wiedererkennbarkeit von Melodien thematisiert und das Spielen mit den Erwartungen der Rezipienten. Außerdem weisen Juslin und Timmers (2010) darauf hin, dass von der ohnehin wenigen Forschung zu musikalisch induzierten, also nicht lediglich erkannten, sondern selbst gefühlten Emotionen praktisch nichts in einem realistischen Setting durchgeführt wurde, sondern überwiegend unter Laborbedingungen mit Beschränkung auf Audio-Aufnahmen.

Kopiez (2005) macht in seinem am Anfang dieses Abschnitts bereits erwähnten Überblickartikel folgende Prognose, die er drei Jahre später nochmals fast wortgleich wiederholt (vgl. Kopiez, 2008, S. 333), welche Themenbereiche innerhalb der Performance-Forschung in nächster Zeit an Bedeutung gewinnen könnten:

„Obwohl man über die weitere Entwicklung nur spekulieren kann, wird vermutlich die experimentelle Interpretationsforschung zukünftig einen starken Akzent auf folgende Forschungsthemen legen: Interpretations-

analyse in Echtzeitverfahren, Visualisierung von Interpretation und ihre pädagogische Anwendung, Weiterentwicklung einer umfassenden empirisch fundierten Interpretationstheorie und die Integration von Forschungen zur emotionalen Wirkung von Musik [...]“ (Kopiez, 2005, S. 513)

Dabei ist nicht ganz klar, ob Kopiez (2005) bzw. Kopiez (2008) mit „Interpretationsanalyse in Echtzeitverfahren“ ausschließlich die computerbasierte Analyse der Audiospur oder des Videos meint, oder auch eine empirische kontinuierliche Untersuchung der Rezeption mit inbegriffen sein könnte. Allerdings ist im Bereich kontinuierlicher Performancebewertung auch zehn Jahre nach der Erwähnung im Forschungsausblick wenig passiert ist und speziell empirische Echtzeitverfahren kamen in diesem Kontext fast nie zum Einsatz.

2.3 Theorien zur Performancebewertung

In dieser Arbeit werden Bewertungsunterschiede für musikalische Darbietungen untersucht und insofern ist es wichtig, Theorien zu betrachten, wie eine solche Bewertung überhaupt entsteht. Da Theoriebildung bzw. die Entscheidung für oder gegen eine Theorie allerdings nicht konkret Teil der Fragestellung ist, wird die Darstellung dieser Theorien hier kurz gehalten. Ein ausführlicherer Überblick findet sich bei Platz (2014), der als Grundlage für die Zusammenfassung in diesem Abschnitt verwendet wird.

Als drei bestehende Erklärungsansätze für die Entstehung einer Bewertung führt Platz (2014) einen ästhetischen Vergegenwärtigungsprozess, einen Kommunikationsprozess und einen psychologischen Bewertungsprozess an. Dabei beziehen sich die ersten beiden Modelle schwächer auf eine konkrete Bewertung. Bei einem ästhetischen Vergegenwärtigungsprozess wird die musikalische Darbietung überwiegend auf eine hermeneutische Auseinandersetzung mit dem Werk reduziert, wobei der interpretatorische Aspekt unberücksichtigt bleibt (vgl. Platz, 2014, S. 6-9). Setzt man voraus, dass die Zuhörer in der Lage sind, die Leistung des Musikers zu bewerten und nicht eine Bewertung des Stücks abgeben, ist dieses Modell für die Bewertungsfrage irrelevant. Es wird scheinbar nicht einmal die Frage gestellt, ob die Performance beispielsweise die Strukturen der Musik, mit denen sich im ästhetischen Vergegenwärtigungsprozess auseinandergesetzt wird, ausreichend darstellt.

Die Kommunikationsmodelle bzgl. musikalischer Performances sind an allgemeine Kommunikationsmodelle angelehnt. Denen zufolge wird die musikalische Darbietung im Sinne eines Kommunikationsprozesses verstanden, bei dem es einen Kommunikator, eine Nachricht und einen Rezipienten gibt. Dabei kodiert der Musiker eine Nachricht, beispielsweise eine zu vermittelnde Emotion, und der Zuhörer dekodiert diese aus der Musik und die Übereinstimmung kann als Signalqualität verstanden werden (vgl. Platz, 2014, S. 9-11; Juslin & Timmers, 2010). Diese Theorie bezieht sich nicht direkt auf einen Evaluationsprozess, allerdings räumt sie dem Performer eine deutlich wichtigere Rolle ein, sodass der Rezipient in die Lage versetzt wird, auch seine Leistung und nicht ausschließlich das Stück zu bewerten.

Psychologische Erklärungsansätze verstehen die Bewertung einer musikalischen Darbietung als allgemeinen Evaluationsprozess, bei dem auch rezipientenseitige Kriterien einen Einfluss haben (s. Platz, 2014, S. 14-16). Diese Herangehensweise ist im Kontext dieser Arbeit die relevanteste, da sie sich konkret auf die Bewertungsprozesse bezieht und rezipientenseitige Faktoren berücksichtigt. Platz (2014) bezieht sich bei dieser Theorie stark auf den Aspekt der Zeitvariabilität, der in den Abschnitten 2.5 und 2.6 noch ausführlicher dargestellt wird, und betrachtet dabei primär den Zeitpunkt des ersten Eindrucks. Er untersucht, wie weit dieser Eindruck vor Beginn des ersten Tons liegen kann, und dessen Bedeutung für den Anfang der Evaluationsphase, aber nicht dessen längerfristigen Einfluss auf ein Gesamturteil.

Ein etwas konkreteres Modell eines psychologischen Bewertungsprozesses, bei dem die Faktoren, die neben der eigentlich zu bewertenden musikalischen Performance einen Einfluss auf die Wertung ausüben, in die Kategorien extra-musikalisch, nicht-musikalisch und Messfehler eingeordnet werden, präsentieren McPherson und Schubert (2004) (s. Abb. 2.1). Dabei sei es für die Juroren unmöglich, ausschließlich die musikalischen Faktoren, wie Technik, Interpretation, Ausdruck und Kommunikation zu berücksichtigen und die Faktoren, die in der Abbildung gestrichelt umrandet sind, von ihrer Bewertung auszuschließen. Messfehler beispielsweise lassen sich ggf. minimieren, aber nicht völlig ausschließen, wobei in diversen Studien davon ausgegangen wird, dass professionelle Musiker mit einer hohen Expertise in der Bewertungssituation akkurater bzw. reliabler bewerten oder dies überprüft und diskutiert wird (vgl. Bergee, 1993; Thompson & Williamon, 2003; Smith, 2004). Die extra-musikalischen Faktoren unterteilen McPherson und Schubert (2004) in performerbezogene Aspekte, wie Selbstwirksamkeit, Abweichung von der Ausdrucksnorm, also dem üblichen Grad an Agogik etc., Attraktivität und Bewegung, kontextbezogene Aspekte, wie Akustik oder die Reaktion des Publikums, und solche

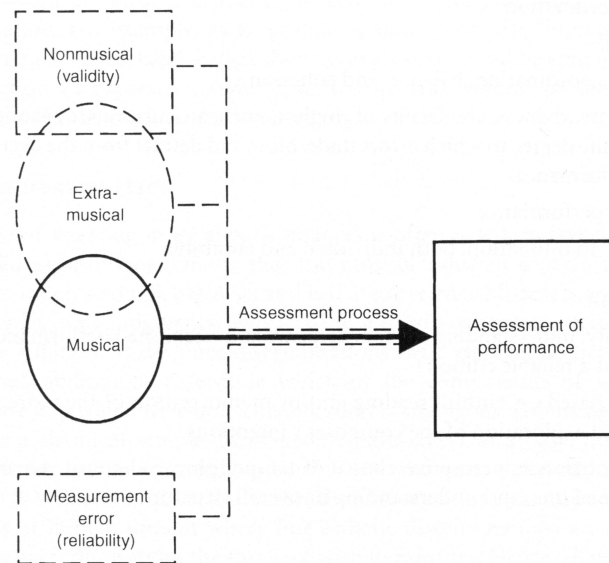


Abbildung 2.1: Modell des Musikbewertungsprozesses
(McPherson & Schubert, 2004, S. 63)

Aspekte, die vom Juror abhängen. Darunter fallen gestellte Erwartungen aufgrund vorigen Wissens über den Musiker, der erste Eindruck, die Stimmung des Jurors, Bekanntheitsgrad des Stückes und Gefallen daran, sowie ein möglicher Halo-Effekt. Bei letzterem wird der Juror stark durch einen einzelnen Aspekt der Performance, den er oder sie besonders wichtig findet, beeinflusst und vernachlässigt die übrigen Aspekte. Unter den nicht-musikalischen Faktoren werden diejenigen gefasst, die im Zusammenhang mit der Validität der Bewertung stehen. Dazu zählen beispielsweise auf Geschlecht oder Hautfarbe basierende Stereotype oder die Reihenfolge der Auftritte.

Einige Zuordnungen zu den groben Kategorien wirken willkürlich, etwa dass die Geschlechterstereotype als nicht-musikalisch gesehen werden, Attraktivität des Performers allerdings als extra-musikalisch, wobei auch hier Stereotype beeinflussen können. Insgesamt jedoch ergibt sich ein umfassendes Bild, welche Arten ganz oder teilweise außermusikalischer Einflussgrößen im Evaluationsprozess eine Rolle spielen können, und, dass sich viele dieser Einflüsse nicht komplett verhindern lassen. Interessant ist, dass nur die als nicht-musikalisch klassifizierten Einflüsse als Einschränkung der Validität gesehen werden, nicht aber die extra-musikalischen, speziell die auf die Juroren bezogenen Faktoren, wie beispielsweise deren Stimmung.

Eine Verzerrung durch Einflüsse des Bewertenden wird also als Teil des Prozesses akzeptiert und nicht ausschließlich als Störfaktor gesehen.

Darüber hinaus stellt Platz (2014) zwei weitere Theorien aus anderen Bereichen vor, die für seine Arbeit besonders relevant sind. Er überträgt sie auf die musikalische Performance, wobei die Modelle nicht gänzlich getrennt betrachtet, sondern miteinander verzahnt werden. Die soziale Interaktionstheorie nach Goffman geht über den Kommunikationsansatz insofern hinaus, dass sie die Darbietung nicht als Informationsübertragungspotenzial versteht, sondern als soziale Handlung, also wechselseitige Beeinflussung, bzw. als Interaktionspotenzial (Platz, 2014, S. 23). Dabei wird die Vorderbühne als öffentlicher Raum verstanden, auf dem der Musiker eine bestimmte soziale Rolle einnimmt. Der Musiker inszeniert sich hier selbst bzw. seine Leistung und versucht, die von ihm erwartete Rolle auszufüllen, was allgemeiner auch als *impression management* im Sinne des Versuchs, dem Gegenüber „einen spezifischen Eindruck von sich selbst zu vermitteln“ (American Psychological Association, 2007, Übersetzung und zit. nach Platz, 2014) bezeichnet wird. Die Hinterbühne hingegen wird als privater Raum betrachtet, in dem der Musiker sich nicht darstellen muss. Der Rollenwechsel findet folglich bei dem Betreten bzw. Verlassen der Vorderbühne statt und in sofern ist das Bühnenauftrittsverhalten die erste Gelegenheit, das Impression Management des Musikers zu beobachten. Auf der Rezipientenseite findet parallel dazu die *Impression Formation* statt, was sich wörtlich zu „Eindrucksbildung“ oder Urteils- bzw. Meinungsbildung übersetzen lässt. Direkt darunter verstanden wird Platz' (2014) Darstellung zufolge die Bildung einer „mentalen Repräsentation“. Dies entspricht unabhängig von dem genau verwendeten Begriff dem psychologischen Bewertungsprozess, der vorab schon als für sich stehende Theorie dargestellt wurde, und hier als Teil eines komplexeren Rahmens gesehen wird. Die erste Urteilsfindung auf Seiten der Zuhörer findet diesem Modell zufolge direkt nachdem der Musiker die Bühne betritt statt und das Ergebnis beeinflusst möglicherweise den weiteren Urteilsbildungsprozess während der Darbietung bzw. das Urteil für die gesamte Performance. Laut Platz (2014) müsse nach der sozialen Interaktionstheorie eine Verletzung der Rollenerwartung, also eine geringe Übereinstimmung des erwarteten Rollenverhaltens mit dem tatsächlich beobachteten, bei den Zuhörern ein negatives Angemessenheitsurteil auslösen. Dieses führe dann zu einer rezipientenseitigen Interaktionsunterbrechung, die wiederum eine schlechtere Bewertung nach sich ziehe. Platz (2014) fand nur eine Studie, die den Einfluss des Angemessenheitsurteils auf die Bewertung insgesamt untersuchte, bei der tatsächlich eine als weniger angemessen bewertete Konzertkleidung bei davon

abgesehen immer gleichbleibenden Versuchsbedingungen auch zu einer schlechteren Beurteilung der Performance führte (vgl. Griffiths, 2008).

Im zweiten Schritt erläutert und überträgt Platz (2014) die persuasive Rhetoriktheorie nach Knappe auf die Situation einer musikalischen Performance und bezieht dabei die soziale Interaktionstheorie weiterhin mit ein, sodass sich ein Gesamtbild ergibt. Die persuasive Rhetoriktheorie basiert auf der Annahme, alle rhetorischen Handlungen, also im Kontext von Musik alle Handlungen eines Musikers auf der Bühne, haben als Ziel die *Persuasion*: die Urteils- bzw. Einstellungsänderung des Rezipienten. Entsprechend ist jegliches Bühnenverhalten, also spieltechnische oder auch davon unabhängige Bewegungen, sowie Mimik, Blickkontakt u. v. m., vom Betreten bis zum Verlassen der Bühne genauso als persuasive Handlung zu verstehen, wie die gespielte Musik selbst. Die Verwendung des Begriffs *Persuasion* bleibt etwas unklar, da Platz (2014) ihn sowohl, wie gerade definiert, als Einstellungsänderung benutzt, allerdings auch schreibt, er bezeichne den „gesamten Überzeugungsvorgang des Interpreten zum Zweck des gezielten mentalen Wechsels des Rezipienten“ (Platz, 2014, S. 29), wobei der mentale Wechsel wiederum der Urteils- oder Einstellungsänderung entspricht. Die Einstellung bzw. Interaktion, die bei den Zuhörern erzielt werden soll, ist die *musikbezogene Performance-Elaboration*, die Platz (2014) als „denkende Auseinandersetzung mit Eindrucksaspekten innerhalb eines musikbezogenen Persuasionsprozesses“ (S. 35) definiert. Die *Performance-Elaboration* ist abhängig von der Motivation und Fähigkeit des Rezipienten, der Performance zu folgen (S. 36). Dabei haben die Impression-Management-Strategien, die zur Persuasion des Zuhörers eingesetzt werden, keine universelle Wirkung sondern eine individuelle, rezipientenspezifische (Platz, 2014, S. 32), ganz im Sinne des Interaktionsmodells, das auf wechselseitiger Beeinflussung basiert. Die persuasive Rhetoriktheorie betrachtet darüber hinaus auch Verarbeitungsweisen des Gesehenen, also ob eine rhetorische Handlung beispielsweise direkt in ein bekanntes Schema eingeordnet werden kann, oder mehr Denkleistung erforderlich ist. An dieser Stelle wird auf eine detailliertere Erläuterung verzichtet, da die exakten Mechanismen hier nicht benötigt werden, um die Theorie anzuwenden.

Bezieht man die Rhetoriktheorie mit ein, wird auch noch die bereits für die Interaktionstheorie dargestellte Wirkungsfolge, auf die Platz (2014) am Ende des Abschnitts zur Rhetoriktheorie ebenfalls zurückkommt, schlüssiger: Wie oben beschrieben führt demnach eine Verletzung der erwarteten Rollendarstellung zu einem negativen Angemessenheitsurteil. Im nächsten Schritt geht es nun um Performance-

Elaboration. Lässt sich der Zuhörer auf die Darbietung ein? Ist das Angemessenheitsurteil positiv und der Rezipient kann den Musiker bzw. sein Impression Management in ein bekanntes Schema einordnen, wird seine Bereitschaft zur Performance-Elaboration höher sein, als wenn er das Rollenverhalten unangemessen findet. Insbesondere ist bei einem Verstoß gegen diese Konventionen die Einordnung in ein Schema nicht problemlos möglich, sondern der Rezipient muss sich basierend auf einzelnen Attributen ein neues Bild machen. Damit lässt sich die vorher bereits benannte Interaktionsunterbrechung erklären. Muss der Zuhörer erst überlegen, was er beispielsweise von der unkonventionellen Kleidung des Musikers hält, verliert er einen Moment die übrigen Aspekte der Performance aus dem Blick. Die Folgerung, dass wenn ein Zuhörer sich nicht auf die Performance einlässt, sich nicht aktiv mit ihr auseinandersetzt, seine Bereitschaft zur Performance-Elaboration also niedrig ist, legt die Vermutung nahe, dass dies seine Bewertung der Performance negativ beeinflusst. Letzteres bleibt allerdings unbelegt, abgesehen von der Studie zur Konzertkleidung von Griffiths (2008), da Platz (2014) das Bühnenauftrittsverhalten in den Mittelpunkt stellt und im Zusammenhang zu dem erhobenen Angemessenheitsurteil ausschließlich die Motivation zum Weiterhören untersucht. Eine globale Bewertung der Performances wurde allerdings nicht erhoben, sodass keine Aussage zum Einfluss des ersten Eindrucks auf das Gesamturteil getroffen werden kann.

Platz (2014, S. 37) setzt diese Theorie durchaus ins Verhältnis zum Modell von McPherson und Schubert (2004), indem er den Aspekt des ersten Eindrucks als „Leistung der rezipientenseitigen, evaluativen Personenwahrnehmung und weniger [als] das Abbild objektiver Interpreteneigenschaften“ auslegt. Über diese benannte Übereinstimmung hinaus gibt es noch eine größere Schnittmenge zwischen den Modellen. Die Stimmung des Juroren nach McPherson und Schubert (2004) steht vermutlich in engem Zusammenhang mit der Motivation zur Performance-Elaboration bei Platz (2014) und die Selbstwirksamkeit des Musikers ist in dieser Situation Teil des Impression Managements in der Performer-Rolle.

Insgesamt klingt in Platz' (2014) Darstellung durchaus an, dass eine Betrachtung der Zeitkomponente entscheidend ist. Bei der Auswertung der Studie von Thompson, Williamon und Valentine (2007) wird der Aspekt explizit genannt und außerdem betrachtet er in einer der beiden Studien explizit das Bühnenauftrittsverhalten als Zeitpunkt der ersten Urteilsfindung sowie dessen Bewertung und die Motivation der Rezipienten, weiter zuzuhören (vgl. Platz & Kopiez, 2013; Platz, 2014, Kap. 6). Er stellt die These auf, dass dieser erste Eindruck bzw. das Angemessenheitsurteil wesentlichen Einfluss auf die Bewertung des weiteren

Verlaufs des Stücks habe, auch wenn er dies im empirischen Teil der Arbeit nicht belegt. Wie Meinungsbildungsprozesse während des Zuhörens und Beobachtens funktionieren bzw. funktionieren könnten, wird an dieser Stelle, wo es mehr um allgemeine Mechanismen geht, nicht weiter ausgeführt, sondern in Abschnitt 2.6 wieder aufgegriffen und deutlich ausführlicher betrachtet.

2.4 Empirische Befunde zu Performancebewertung

2.4.1 Visuelle Einflüsse auf die Performancebewertung

Es gibt bereits zahlreiche Studien die den Einfluss bestimmter visueller Einflüsse auf die Bewertung musikalischer Darbietungen messen oder allgemein die Bewertungen einer audiovisuellen und ausschließlich gehörten Präsentationsform einer Performance vergleichen. Huang und Krumhansl (2011) ließen einen Pianisten drei verschiedene Stücke spielen, jeweils eine Variante mit minimalem, normalen und übertriebenen Bühnenverhalten und präsentierten diese ihren Probanden sowohl als Audioaufnahme als auch als Video. Die Teilnehmer, die entsprechend ihrer musikalischen Vorbildung in zwei Gruppen aufgeteilt wurden, bewerteten diese Performances dann auf diversen verschiedenen Skalen. Ein allgemeiner Effekt wurde weder für die Präsentationsform noch für musikalische Expertise der Probanden gefunden, allerdings für das gespielte Klavierstück. Welcher Grad an Bühnenverhalten bevorzugt wurde, war abhängig vom gespielten Stück, wobei die minimalistische Bedingung immer am schlechtesten bewertet wurde, während der Unterschied zwischen normalen und übertriebenen Bewegungen eher gering war. Besonders interessant ist, dass Musiker auch die Audio-Aufnahmen je nach Bühnenverhalten, dass sie in dem Fall nicht sehen konnten, unterschiedlich bewerteten, woraus die Autoren schließen, dass das Bühnenverhalten signifikanten Einfluss auf die gehörte Musik und somit die Einschätzung deren Qualität hat. Die Unterschiede zwischen minimalem, normalem und übertriebenem Bühnenverhalten wurden bei den Versuchspersonen mit musikalischer Vorbildung in der audiovisuellen Bedingung nochmals verstärkt, sodass trotz der Erkennbarkeit in der reinen Audio-Fassung, das Visuelle eine wichtige Rolle spielte. Nicht-Musiker konnten die drei Grade des Bühnenverhaltens nur in der audiovisuellen Bedingung unterscheiden und selbst da waren die Wertungsunterschiede geringer als bei den Personen mit musikalischer Expertise. (vgl. Huang & Krumhansl, 2011). Dass Musiker generell feiner zwischen verschiedenen Performances unterscheiden können, war zu erwarten, allerdings ist interessant, dass Nicht-Musiker offenbar stärker vom Visuellen abhängig sind,

um Unterschiede wahrzunehmen. Dadurch ist jedoch nur gezeigt, dass es einen Unterschied macht, den Musiker zu sehen, nicht aber, ob die gleiche Musik durch beispielsweise mehr Bewegungen besser wahrgenommen wird.

Griffiths (2008) hat dieses Problem insofern umgangen, dass sie in ihrer Studie zwar verschiedene Geigerinnen in unterschiedlicher Konzertkleidung spielen ließ, allerdings die gleiche Aufnahme eines anderen Geigers darüber gespielt hat. Somit sind eventuelle Bewertungsunterschiede tatsächlich auf die visuellen Aspekte zurückführbar. Die Zuhörer favorisierten eindeutig ein langes schwarzes Kleid mit langen Ärmeln gegenüber einem kurzen ärmellosen Kleid (beschrieben als „Nightclubbing dress“), sowie Jeans und T-Shirt, indem sie die Angemessenheit signifikant besser und nur für das lange Kleid überhaupt positiv bewerteten. Teilweise beeinflusste die Konzertkleidung auch die Bewertung, sodass das technische Können einer Geigerin im Konzertkleid bedeutend besser bewertet wurde als im Nachtclub-Kleid. Dies wurde allerdings immer nur in Interaktion mit der Musikerin betrachtet. Daher ist zu vermuten, dass nicht überprüft wurde, inwiefern es einen globalen Effekt der Konzertkleidung für die Bewertung des technischen Könnens oder der Musikalität der Performance gab, oder dieser nicht signifikant wurde und daraufhin nicht im Artikel dokumentiert wurde. Bei der statistischen Analyse wurden auch lediglich die Versuchsbedingungen als unabhängige Variable untersucht, allerdings nicht direkt der Einfluss des Attraktivitätsurteils oder des Urteils der Angemessenheit der Kleidung auf die Bewertung des technischen Könnens oder der Musikalität. Da der Effekt der Bedingung auf die Bewertung der Performance betrachtet wurde und das Angemessenheitsurteil der Kleidung einen starken Zusammenhang mit der tatsächlichen Kleidung hatte, wären die Ergebnisse dafür vermutlich ähnlich.

Neben dem Bühnenverhalten und der Kleidung wurden viele weitere mögliche visuelle Einflüsse systematisch untersucht. Wapnick und seine Arbeitsgruppe verglichen in diversen Studien die Bewertungen für audiovisuelle und reine Audio-Bedingung für alle denkbaren Niveaustufen von Schülervorspiel bis hin zu professionellen Musikern, lediglich beschränkt auf westlich-klassische Kunstmusik (vgl. Wapnick, Darrow, Kovacs und Dalrymple, 1997; Wapnick, Darrow und Mazza, 1998; Wapnick, Mazza und Darrow, 2000; Ryan, Wapnick, Lacaille und Darrow, 2006; Wapnick, Campbell, Siddell-Strebel und Darrow, 2009). Bzgl. des Einflusses von Attraktivität kamen sie beispielsweise zu unterschiedlichen Ergebnissen in den einzelnen Studien, so dass teilweise gutes Aussehen für die Bewertung von Vorteil war, teilweise aber auch keinen Effekt hatte. Darüber hinaus wurden u. a. auch die visuellen Aspekte Kleidung und Bühnenverhalten betrachtet.

M. Lehmann und Kopiez (2011) betrachten im Gegensatz zu praktisch allen anderen Studien keine klassische Musik, sondern ließen Rock-Gitarristen bewerten. Dabei ging es vorwiegend um den Einfluss des eingeschätzten Schwierigkeitsgrades, allerdings stellte sich heraus, dass auch Musikstudierende diesen Schwierigkeitsgrad nicht (besser) einschätzen konnten, als Probanden mit weniger musikalischer Expertise. Darüber hinaus stellen sie aber auch heraus, dass einige der Bewegungen von Gitarristen auf der Bühne sich spieltechnisch gar nicht hilfreich, sondern eher erschwerend auswirken, und ausschließlich als Show für das Publikum eingesetzt werden (M. Lehmann & Kopiez, 2011, S. 199).

Auch Peddell (2008) untersuchte u. A. den Bewertungsunterschied zwischen audio bzw. audiovisueller Präsentation aus verschiedenen Blickperspektiven, wobei immer der Dirigent im Zentrum des Bildes war, aber erhob darüber hinaus Kommentare, worauf die Probanden geachtet haben. Der Fokus liegt hier stark auf dem Dirigentenverhalten und als Bewertung der gesamten Performance wurde der Mittelwert von kontinuierlich erhobenen Daten benutzt. Eine abschließende Gesamtbewertung wurde in der Studie nicht abgefragt, eine tatsächliche Auswertung der Zeitreihen aus den Echtzeitdaten findet allerdings auch weder in diesem Artikel statt, noch in der Dissertation, in der die Daten aus der Studie noch ausführlicher analysiert werden (vgl. Peddell, 2004). Peddell (2008) kommt zu dem Ergebnis, dass die reine Audio-Bedingung im Durchschnitt über die Hördauer besser bewertet wird, als die Video-Präsentation, unabhängig von der Beobachterperspektive. Dieses Ergebnis ist relativ erstaunlich, weil in den meisten Studien die audiovisuelle Bedingung zu besseren Bewertungen führt.

Platz und Kopiez (2012) untersuchten dies näher in einer Meta-Analyse, aus der die Ergebnisse von Peddell (2004) allerdings aufgrund der kontinuierlichen Datenerhebung ausgeschlossen wurden. Berücksichtigt wurden hingegen 17 Studien mit Skalen für Gefallen (engl. „liking“), Ausdruckskraft (engl. „expressiveness“) und die Performance-Qualität, für die alle benötigten deskriptiven Daten angegeben waren – darunter fünf Studien von Wapnick und seiner Arbeitsgruppe, von denen vier oben bereits aufgeführt wurden. Die Meta-Analyse schätzte die mittlere Effektstärke für den Einfluss der visuellen Komponente auf die Gesamtbewertung auf $d = .51$, was einem mittleren Effekt entspricht, wobei zur Berechnung die Bewertungen der reinen Audio-Aufnahme von denen der audiovisuellen Präsentation abgezogen wurde. Die Probanden bewerten demnach im Mittel audiovisuelle Performances um eine halbe Standardabweichung besser als die dazugehörige Audioaufnahme ohne Bild, wobei die Effektgrößen für die einzelnen einbezogenen Studien von $d = .35$ bis

.92 reichten, also bei keiner die audiovisuelle Bedingung durchschnittlich schlechter bewertet wurde.

Auch wenn die Ergebnisse von Peddell (2008) denen von Platz und Kopiez (2012) widersprechen, zeichnet sich hier ein recht eindeutiges Bild, dass die visuelle Komponente einen entscheidenden Teil der Performance bzw. deren Bewertung ausmacht, insofern es nicht sinnvoll ist, Performancebewertung anhand von Audio-Aufnahmen zu untersuchen. Solche Ergebnisse wären folglich nur eingeschränkt aussagekräftig für eine reale Konzertsituation. In wie weit es einen vergleichbar großen Unterschied zwischen der Bewertung von auf Video aufgezeichneten Performances und echten Live-Auftritten gibt, ist aufgrund der aktueller Literaturlage nicht zu beurteilen. Das liegt einerseits daran, dass es nur wenige Studien gibt, die Performancebewertung Live untersuchen, andererseits, dass sich für Live-Situationen die Intra-Rater-Reliabilität, also die Wiederholbarkeit einer Bewertung eines Raters, die für eine Meta-Analyse wie der von Platz und Kopiez (2012) benötigt wird, nicht experimentell ermitteln lässt, da die Live-Situation an sich nicht wiederholbar ist.

2.4.2 Konsistenz von Performancebewertungen

Es sind nur wenige Studien publiziert, die die Konsistenz von Performancebewertungen untersuchen und z. T. wird der Konsistenz-Begriff auch anders ausgelegt, als in dieser Arbeit. Wapnick, Jasinskas, Flowers und Alegant (1993) untersuchen die intraindividuelle Konsistenz von Klavier-Performance-Evaluationen mit einem Versuchsdesign, bei dem immer zwei verschiedene Performances gehört wurden, und dann entschieden werden sollte, welche die bessere sei. Dabei wurde die Bewertung dreier Stücke als konsistent angesehen, wenn nachdem bereits im direkten Vergleich das erste Stück besser als das zweite und das zweite besser als das dritte bewertet wurde, auch anschließend das erste Stück besser als das dritte bewertet wurde. Insofern wurde untersucht, ob die Zuhörer eine in sich konsistente Rangfolge der gehörten Stücke aufstellen konnten. Wenn dies sicher gelingt, kann vermutet werden, dass die Stücke bei wiederholtem Hören wieder vergleichbar gut eingeschätzt werden. Aufgrund der anderen Auslegung des Begriffs, sind die Ergebnisse allerdings für diese Arbeit nicht relevant.

Bergee (1993) untersuchte die Übereinstimmung der Bewertungen von Performances von Blechbläsern mit Korrelationen. In seinen drei Durchgängen des Experiments bewertete je eine Expertenjury aus Dozenten Live-Performances von Studierenden und anschließend die Studierenden der Peer-Group, die Videoaufzeichnungen, wobei auch jeder seinen eigenen Auftritt bewertete. Die Inter-Rater-Reliabilität war

dabei generell hoch, gemessen an den Korrelationen von .83 bis .89 zwischen je zwei Bewertenden und auch die Korrelation zwischen den Gesamtwertungen der Dozentenjury und der Peer-Group war in jedem Setting mit $.83 \leq r \leq .91$ hoch. Lediglich die Selbstevaluation korrelierte schlecht sowohl mit den Bewertungen der Dozenten als auch der übrigen Studierenden. Die Videoaufnahmen wirkten sich gegenüber der Live-Performance nicht beeinträchtigend auf die Bewertungen aus, nicht einmal für die Skala „Tonqualität“.

Davidson und Coimbra (2001) führten ebenfalls eine Studie mit Musikstudierenden in einer Live-Situation durch. Sie untersuchten die Bewertung der Zwischenprüfungen der Gesangsstudierenden im zweiten Studienjahr an der Guildhall School in London und betrachteten dabei einerseits quantitative Daten in Form von Noten und führten andererseits eine qualitative Analyse der Bewertungsgespräche durch. Die Jury bestand dabei aus zwei Gesangslehrerinnen, die z. T. auch die zu prüfenden Studierenden unterrichteten, dem „Head of Department“ für Gesang und einem externen Juror. Als Notenskala wurden die allgemein in Großbritannien üblichen „percentage grades and degree classifications“ (Davidson & Coimbra, 2001, S. 37) benutzt, obwohl die Guildhall School normalerweise Noten von A bis E verwendet, um mit einer Intervallskala rechnen zu können. Die Notenskala reicht theoretisch von 0 % bis 100 %, wobei der Bereich i. d. R. nicht vollständig ausgenutzt wird, und jeder Abschnitt eine Entsprechung wie „first class honours“ oder „second class honours – first/second grade“ hat. Bei der Auswertung fiel auf, dass der externe Juror insgesamt besser bewertete und seine Noten weniger variierten. Insgesamt war die Inter-Rater-Korrelation zwischen zwei Juroren in den meisten Fällen mittel bis hoch ($.6 \leq r \leq .71$), allerdings betrug eine Korrelation lediglich $r = .4$ und war nicht signifikant und eine weitere betrug nur $r = .5$. In einigen Fällen passten die Dozenten ihre Note bei der Diskussion nach jeder einzelnen Prüfung noch an und aus dem Artikel geht nicht klar hervor, ob für die Berechnung die ursprünglichen oder ggf. auch bereits die veränderten Noten verwendet wurden. In allen Fällen wurde bei den als Mittelwert der vier Einzelnoten gebildeten Gesamtnote nur ein Notenspektrum von „second class honours – grade one“ bis „second class honours – grade two“ ausgereizt, welches laut den Autorinnen ein typisches Notenspektrum ist, obgleich es verwunderlich ist, dass kein Studierender als gut genug gesehen wurde um eine „first class honours“ zu bekommen, aber auch niemand schlechter bzw. in einem Fall bei einer schlechter wahrgenommenen Prüfung dennoch beschlossen wurde, eine 2.2 zu geben. Innerhalb der Bewertungsgespräche sei es praktisch nicht zu grundsätzlichen Meinungsverschiedenheiten über die Leistung gekommen, sondern

eher zu verschiedenen Schwerpunkten, welche Einzelaspekte der Darbietung wie gewichtet werden sollten. Darüber hinaus werden anhand der qualitativen Daten die Bewertungskriterien analysiert, welche im folgenden Abschnitt genauer erläutert werden.

Thompson und Williamon (2003) führten eine ähnliche Studie durch, für die Performances von Musikstudierenden verschiedener Instrumentengruppen des Royal College of Music in London auf Video aufgezeichnet wurden, die von externen, erfahrenen Juroren, die die Studierenden nicht kannten, bewertet wurden. Die Korrelationen zwischen den Bewertungen, die für verschiedene Skalen, darunter auch Gesamtqualität, sind mit einer mittleren Korrelation von $\rho = .498$ und einer Range von $.332 \leq \rho \leq .651$ nur moderat. Dadurch, dass hier ein anderer Korrelationskoeffizient gewählt wurde, sind die Ergebnisse allerdings nicht direkt mit denen aus den übrigen Studien vergleichbar. Interessant an dieser Studie ist bzgl. der Inter-Rater-Reliabilität, dass es Anzeichen gibt, dass das Instrument des Raters einen Einfluss auf die Bewertung hat. Ein Streicher bewertet konsequent alle Darbietungen auf Streichinstrumenten etwas schlechter als die anderen Juroren, bei welchen diese Auffälligkeit bei der eigenen Instrumentengruppe hingegen nicht auftritt.

Es gibt auch eine Untersuchung von Smith (2004) in einem professionelleren Kontext, nämlich bei der Vorrunde eines internationalen Streicherwettbewerbs, allerdings wurde diese Wettbewerbsrunde nicht in einer tatsächlichen Auftrittssituation, sondern lediglich anhand von Audioaufnahmen durchgeführt. Dafür hörten sich fünf renommierte Experten gemeinsam die Aufnahmen an, konnten anschließend darüber diskutieren und mussten jeder eine Wertung auf einer fünfstufigen Skala von „nicht qualifiziert (0)“ bis „qualifiziert ohne Einschränkungen (4)“ abgeben. Dabei wurde für jede Stufe eine genaue Beschreibung angegeben, was diese Punktzahl bedeuten solle. Smith (2004) argumentiert, dass alle ihre Bewertungen bereits vor der Diskussion gemacht hätten, und insofern die vergebenen Punktzahlen auf Wertungsunterschiede hin ausgewertet werden könnten. Allerdings bleibt zu berücksichtigen, dass bereits vorher nonverbal kommuniziert worden sein kann, speziell Begeisterung für eine Aufnahme oder Ablehnung könnten anhand von Mimik und Körperhaltung zu großen Teilen erkennbar sein. Die Juroren benutzen von sich aus auch halbe Punkte, woraus Smith (2004) schließt, dass eine fünfstufige Skala für diese Bewertung nicht ausreichend sei. Die Autoren wenden drei verschiedene Maße für die Einigkeit der Juroren an: Einerseits zählen sie, für wie viele der gehörten Bewerber Meinungsverschiedenheiten vorlagen, wobei diese, wenn sie sie

als Abweichung um mind. zwei ganze Skalenpunkte innerhalb der Gruppe definierten, in 36 % der Fälle auftraten. Definiert als Abweichung um mehr als einen vollen Skalenpunkt, gab es in 63 % der Fälle Meinungsverschiedenheiten. Darüber hinaus sind die Extrema bzgl. der paarweisen Inter-Rater-Korrelation angegeben mit $.48 \leq r \leq .78$ und es wurde das Cronbach's α für die Inter-Rater-Reliabilität berechnet ($\alpha = .9$).

Im Kontext von kontinuierlicher Performancebewertung kommen Thompson et al. (2007) zu dem Ergebnis, dass die Variabilität der Echtzeitdaten sowohl inter- als auch intraindividuell hoch ist, spezifizieren dies allerdings nicht konkreter, sodass ein Vergleich nicht möglich ist,

Offensichtlich ist Bergee (1993) zu deutlich besseren Ergebnissen gekommen als Davidson und Coimbra (2001) und Smith (2004) in den neueren Studien, aber über die Ursachen kann hier nur spekuliert werden. Es ist zu erwarten, dass bei dem internationalen Streicherwettbewerb ausschließlich sehr gute Aufnahmen eingereicht wurden und, dass es schwieriger war zu unterscheiden, als bei den tatsächlichen Live-Performances, wie Bergee (1993) sie untersucht hat. Diese Erklärung reicht aber nicht, um die noch niedrigeren Übereinstimmungen bei Davidson und Coimbra (2001) zu erklären. Eine weitere mögliche (Teil-)Ursache ist die verwendete Bewertungsskala, auf der bei Davidson und Coimbra (2001) nur der mittlere Bereich ausgenutzt wurde und dadurch bei der Analyse mittels Korrelationen oftmals Unterschiede von wenigen Prozentpunkten ins Gewicht fallen, die aber kaum einen inhaltlichen Unterschied darstellen. Dass die Skala an dieser Stelle allein ausschlaggebend für die unterschiedlich hohen Wertungsübereinstimmungen ist, ist zwar unwahrscheinlich, aber da grundsätzlich davon auszugehen ist, dass sie Einfluss auf die Bewertungen hat, werden die verschiedenen Möglichkeiten im folgenden Abschnitt noch genauer thematisiert. Auch die Überlegung, dass bei Bergee (1993) alle Wertenden selbst Blechbläser waren, kann den Unterschied nicht erklären, da bei Davidson und Coimbra (2001) auch ausschließlich Sänger in der Jury saßen.

2.4.3 Bewertungsskalen

Es gibt grundsätzlich zwei Typen von Bewertungsskalen, die für musikalische Darbietungen eingesetzt werden: Einerseits sind holistische Gesamtbewertungen auf einer einzigen Skala für die Gesamtqualität, Leistung o. Ä. möglich, z. B. auch eine Notenskala, oder alternativ werden auf diversen Skalen viele Einzelaspekte der Performance erfasst, wobei in beiden Fällen eine Einschränkung auf die reine musikalische Darbietung möglich ist oder das auf die visuelle Komponente und ggf.

Auf- und Abgang ausgeweitete, ganzheitliche Verständnis von Performance, wie in Abschnitt 2.1 definiert, verwendet werden kann.

In den im vorigen Abschnitt vorgestellten Studien kommen beide Skalentypen zum Einsatz. Davidson und Coimbra (2001) verwenden eine Notenskala und versuchen außerdem die verwendeten Bewertungskriterien in offenen Diskussionen herauszufinden und Smith (2004) argumentiert, dass sofern es lediglich um ein Ranking gehe, eine einzelne Gesamtskala ohne viele Erklärungen ausreiche. Bergee (1993) hingegen verwendet über 20 Einzelskalen, die in vier verschiedene Bereiche untergliedert werden, und berechnete daraus einen Gesamtscore. Er argumentiert, dass die Ergebnisse nahelegen, dass die Bewertung eher global und unabhängig von einzelnen abgefragten Aspekten vorgenommen wird. Thompson und Williamon (2003) vergleichen ebenfalls verschiedene Skalen, neben der Gesamtqualität auch die drei Bereiche wahrgenommenes instrumentales Können (engl. *perceived instrumental competence*), Musikalität und Kommunikation, wobei für jede Kategorie nochmals eine globale Skala sowie einige einzelne Subskalen abgefragt werden. Eine große Validität erwarten sie eher bei holistischer Gesamtwertung und in ihrer Studie stellte sich heraus, dass die Korrelationen zwischen den einzelnen Kategorien und der Gesamtwertung mit $.86 \leq \rho \leq .975$ höher waren, als die Inter-Rater-Korrelationen. Thompson und Williamon (2003, 35f.) geben vier mögliche Erklärungen dafür an, dass die Wertungen von den Subskalen so stark mit den Gesamtbewertungen zusammenhingen, obgleich sie darauf hinweisen, dass die Korrelationen natürlich nicht bedeuten, dass die Bewertungen tatsächlich ähnlich gut oder schlecht waren. Einerseits könnten die Juroren zunächst zu einer holistischen Gesamtbewertung gekommen sein, und anschließend die Subkategorien ähnlich ausgefüllt haben, sofern die Performance nicht in einer Kategorie offensichtlich davon abwich. Dies entspricht auch der von Bergee (1993) vermuteten Ursache. Die zweite Möglichkeit ist, dass die Juroren nicht in der Lage waren, bedeutungsvoll zwischen den Kategorien zu unterscheiden, ggf. auch weil die Kategorien nicht die waren, die sie üblicherweise anwenden würden. Nach Variante drei können die Juroren sehr wohl zwischen den Kategorien unterscheiden, allerdings weichen sie nicht stark voneinander ab, da es einen kausalen Zusammenhang zwischen ihnen gibt. Zum Schluss ist noch denkbar, dass es lediglich in dieser Studie nicht funktioniert hat, da bei den Darbietungen keine dabei war, bei der die Kategorien stark voneinander abwichen.

Die Ursache kann anhand solcher quantitativer Studien nicht geklärt werden, sondern eher durch eine qualitative Analyse der Bewertungsgespräche wie bei Davidson und Coimbra (2001). Insgesamt bleibt jedoch festzuhalten, dass sowohl die

besten, als auch die schlechtesten Inter-Rater-Korrelationen bei Studien mit diversen Skalen vorliegen und diejenigen mit einem einzigen holistischen Gesamtrating dazwischen, sofern man die Ergebnisse lediglich anhand der Korrelationen vergleicht, ohne explizit die Unterschiede im Versuchsdesign zu berücksichtigen. Insgesamt lässt sich so nicht festlegen, welche Skala einheitlichere Bewertungen begünstigt, aber aus der Erhebung diverser Subskalen ist demnach kein substanzieller inhaltlicher Mehrwert zu gewinnen. Geht es lediglich um eine Bewertung in Form einer Note oder ein Ranking ist somit der Argumentation von Smith (2004), eine einzige Intervallskala sei ausreichend, zuzustimmen.

Einen Spezialfall stellt in diesem Kontext die Studie von Thompson et al. (2007) dar, bei der die Probanden Klavierperformances kontinuierlich während des Hörens bewerteten. Die Teilnehmer bewerteten zwar nach Ende jedes Stückes drei verschiedene Skalen, Gesamtqualität, Musikalität und technisches Können, allerdings mussten sie in dem Echtzeitverfahren in drei Gruppen aufgeteilt werden, sodass jeder nur eine der Skala bewertete. Dabei ergab sich für Gesamtqualität und Musikalität ein ähnlicher Verlauf, bei dem die Bewertungen sich bereits vor Ende des Stückes bei einem Wert einpendelten und kaum mehr veränderten, für technisches Können hingegen wich dieser ab. Da die abschließenden Bewertungen sich nicht signifikant danach unterscheiden, welche Kategorie vorab während des Hörens bewertet wurde, ist daher eher von einer unterschiedlichen Entwicklung der einzelnen Kategorien auszugehen oder einem länger anhalten Fokus auf Details beim Bewerten der technischen Souveränität, als davon, dass der Fokus die Einschätzung der Darbietung insgesamt verzerrt. Allerdings ergibt sich bei diesem Design für jede Bedingung nur eine kleine Stichprobe, sodass eine Verallgemeinerbarkeit der tatsächlichen Bewertung der Performance nur eingeschränkt möglich ist und auch statistische Tests selten signifikant werden (Thompson et al., 2007).

Ein Problem bei der Verwendung von Noten als Skala für die Untersuchung von Performancebewertungen belegen die Studien von Wolf und Kopiez (2014), sowie A. C. Lehmann (2014), die jeweils an verschiedenen Musikhochschulen die Entwicklung von der Eignungsprüfung bis hin zu einer Abschlussprüfung nach 3 bzw. 4 Jahren untersuchen und dabei eine Inflation der Noten feststellen. Bei A. C. Lehmann (2014) sind die durchschnittlichen Noten für das Hauptinstrument sowohl bei Eignungsprüfungen als auch bei den Abschlussprüfungen besser als 1.9. Dadurch dass im instrumentalen Bereich durchgehend sehr gute Noten gegeben werden, entstehen Deckeneffekte, sodass wirklich herausragende Leistungen nicht mehr als solche ausgezeichnet werden können und generell kaum Unterschiede

bzgl. der Leistung daraus abzuleiten sind. Werden also Noten zur Bewertung von Performances eingesetzt, kann dieser Effekt die Ergebnisqualität deutlich beeinträchtigen, da sowohl Dozenten als auch Studierende Noten entsprechend ihrer Erfahrung geben würden.

2.5 Kontinuierliche Datenerhebung

Um die Zeitkomponente, ohne die Musik nicht existieren kann, bei ihrer Wahrnehmung mit zu erfassen, werden sogenannte *continuous response*-Verfahren eingesetzt. Dabei werden die Probanden aufgefordert, während des Musikhörens durchgehend mit Hilfe eines Interfaces eine Bewertung auf einer oder mehreren Skalen abzugeben oder aus einer Auswahl eine Kategorie auszuwählen, und die Antworten werden kontinuierlich, bzw. genauer, zu äquidistanten Zeitpunkten aufgezeichnet. Für diese Erhebungsform hat sich der Begriff *continuous response* durchgesetzt, oder alternativ wird im Englischen geschrieben, dass die Antworten, zumeist Selbsteinschätzungen, in *realtime* aufgezeichnet werden. Entsprechend etabliert haben sich im Deutschen die Übersetzungen *kontinuierlich* bzw. *Echtzeit*, wobei der Teil „response“ nicht wörtlich zu Antwort oder Reaktion übernommen wird, sondern stattdessen meist spezifiziert wird, dass beispielsweise emotionale Selbstauskünfte oder Präferenzurteile erhoben werden, oder es wird direkt der komplette englische Begriff übernommen.

Dabei kritisieren sowohl Schubert (2010) als auch Kopiez, Dressel, Lehmann und Platz (2011), dass dieser Begriff ungünstig gewählt sei, weil eine echte Kontinuität im Zeitverlauf nicht möglich sei. Darüber hinaus weisen beide darauf hin, dass kontinuierlich sich hierbei auf die Zeit beziehe, auch wenn bei diesen Verfahren oft zudem kontinuierliche bzw. sehr feinstufige Skalen verwendet werden. Die Probanden sind tendenziell schon in der Lage, die Frage tatsächlich kontinuierlich zu beantworten, vorausgesetzt sie schaffen es, sich durchgehend auf die Aufgabe zu konzentrieren, was ggf. problematisch ist, wenn sie sich auch auf die Musik konzentrieren sollen. Unmöglich ist allerdings eine kontinuierliche Aufzeichnung der Daten, da diese digital erfolgt, folglich nur eine diskrete Anzahl einzelner Samples in einem Zeitraum gespeichert werden kann. Durch die hohe Frequenz, mit der Daten mit modernen Computern festgehalten werden können, könne dennoch die Illusion kontinuierlicher Bewertungen entstehen (vgl. Schubert, 2010, S. 224). Schubert (2010, S. 225) argumentiert insofern, der Begriff solle immer als „time-dependant response“, also zeitabhängige statt durchgehende Antwort interpretiert werden.

In der Forschung zu Musik und Emotionen sind Echtzeitverfahren bereits relativ etabliert, insbesondere bei der Verwendung eines zweidimensionalen Emotionsraums, dargestellt als Koordinatensystem mit den Dimensionen Valenz und Arousal (Nagel, 2007; Juslin & Timmers, 2010; Schubert, 2010), aber auch eindimensionale Messverfahren mit nur Arousal oder der Emotionsintensität als Dimension kommen zum Einsatz (Madsen, 1998). Darüber hinaus sind kontinuierliche emotionale Selbstauskünfte auch für kategoriale Emotionsmodelle denkbar, wenngleich diese gegenüber dimensional Modellen eine untergeordnete Rolle spielen. Es gibt innerhalb der Musikpsychologie zahlreiche weitere Anwendungsfelder, wie z. B. die Lautstärkenwahrnehmung (Geringer, 1995) oder die Entwicklung von Präferenzurteilen während des Hörens (Brittin & Sheldon, 1995), die zumindest in einzelnen Studien bereits in Echtzeit betrachtet wurden. Bisher existieren nur wenige, explorative Studien, die eine solche Methodik für Performancebewertungen anwenden (Peddell, 2004; Thompson et al., 2007).

Für diese Anwendungsfelder wurden entsprechend passende Interfaces entwickelt, wovon das erste, nach wie vor bekannte, das in den späten 80-ern entstandene CRDI (Continuous Response Digital Interface) ist, das aus einem Zeiger auf einer Drehscheibe oder einem Kästchen mit einem Hebel besteht und somit zeitbezogen kontinuierliche Bewertungen auf einer kontinuierlichen Ratingskala ermöglicht. Modernere Interfaces zur Erhebung von Echtzeitdaten in einer oder zwei Dimensionen funktionieren computerbasiert und werden mit einer Maus oder einem Joystick gesteuert, wobei neben den „offiziellen“ Interfaces, die allgemein verfügbar sind und zu denen es zumeist auch eigene Publikationen gibt, auch Eigenentwicklungen in Studien eingesetzt werden. Einen sehr guten Überblick über die Vielzahl an Interfaces, die für Echtzeitverfahren entwickelt wurden, geben Kopiez et al. (2011), wobei dennoch zu berücksichtigen ist, dass diese wenige Jahre alte Publikation in mancher Hinsicht bereits veraltet ist. Das Problem ist, dass die computerbasierten Interfaces sehr schnell nicht mehr auf dem aktuellen Stand der Plattformen sind, wenn sie nicht konsequent weiter entwickelt werden, sodass selbst die neueren Programme, die dort vorgestellt werden, nicht mehr unbedingt problemlos einsetzbar sind. Beispielsweise die von Nagel (2007) entwickelte, javabasierte Software EMuJoy (Emotion measurement while listening to Music using a Joystick), die neben der etwa zeitgleich entstandenen Software RTCRR (Real Time Cognitive Response Recording) als letztes eigenständiges, nicht internet-basiertes Interface für zweidimensionale Datenerhebung aufgeführt wird (Kopiez et al., 2011, S. 131-149), ist mit einer aktuellen Java-Version nicht mehr bzw. nur sehr schwierig zum Laufen

zu bringen. Relevanter werden zudem mobile Interfaces, wie das bei Kopiez et al. (2011, S) erwähnte pARF (Portable Audience Response Facility), das auf mobilen PDA-Geräten läuft, oder die neuere Entwicklung *emoTouch* für iPad (Louven & Scholle, in Vorbereitung), die auch in Live-Konzertsituationen eingesetzt werden können und somit überhaupt die kontinuierliche Bewertung von Live-Performances ermöglichen.

2.5.1 Sample-Rate und Latenzzeiten

Die empfohlenen Sample-Raten für die Aufzeichnung kontinuierlicher Selbstauskünfte zwischen 2 und 30 Hz orientieren sich zum Einen an der Fragestellung und dem erwarteten Antwortverhalten im Bezug auf die Frequenz mit der Änderungen stattfinden, um alle inhaltlich bedeutsamen Veränderungen zu erfassen, dabei aber möglichst wenig Redundanzen aufzuzeichnen. Auf der anderen Seite hängt die Sample-Rate von der Leistungsfähigkeit der Computer ab, wobei diesem Aspekt immer weniger Bedeutung zukommt. Schubert (2001, 403f.) empfahl noch, emotionale Selbstauskünfte zweimal pro Sekunde zu speichern, allerdings sind inzwischen Sample-Raten von 20-30 Hz üblich bzw. technisch problemlos zu handhaben (Nagel, 2007; Schubert, 2010) und eine Reduktion auf eine geringere Genauigkeit findet nur noch aus inhaltlichen Gründen statt. Die Vorteile liegen dabei, dass einerseits schnellere Veränderungen aufgezeichnet werden können, aber auch darin, dass die Synchronisation mit den eingesetzten Stimuli sowie zusätzlich erhobenen physiologischen Daten wie Hautleitfähigkeit oder Herzfrequenz einfacher wird (Nagel, 2007). Außerdem ermöglicht eine Sample-Rate, die vergleichbar ist mit der Bildfrequenz von Filmen, eine visuelle, bewegte Darstellung der aufgezeichneten Daten in Echtzeit zusammen mit den Stimuli, ohne dass einzelne Datenpunkte sichtbar werden und ohne dass diese durch Zwischenwerte ergänzt werden müssen, beispielsweise durch lineare Interpolation.

Als Latenzzeit oder Reaktionszeit wird die Zeit bezeichnet, die die Probanden benötigen um auf Ereignisse im Stimulus zu reagieren und ihre Reaktion mit Hilfe des Interfaces zu anzugeben. Diese zeitliche Verzögerung wird im Hinblick auf die Wahl einer geeigneten Sample-Rate betrachtet, aber ist auch für die Analyse von Zusammenhang zwischen den eingesetzten Stimuli und den Reaktionen der Probanden entscheidend. Einheitliche, belastbare Angaben bzgl. der zu erwartenden Reaktionszeiten gibt es allerdings kaum. Schubert (2010, S. 226) zufolge finden Reaktionen auf graduelle Lautstärkenveränderungen innerhalb von zwei bis vier Sekunden statt, aber bei plötzlichen lauten Ausbrüchen, ist die Latenzzeit kürzer.

Emotionale Selbstauskünfte könnten sogar innerhalb von weniger als einer Sekunde gemacht werden, die Reaktionszeit sei allerdings auch von der Anzahl der Dimensionen, die gleichzeitig berücksichtigt werden müssen abhängig. Schäfer, Zimmermann und Sedlmeier (2014) gehen von einer Latenz von fünf Sekunden für emotionale Selbstauskünfte im zweidimensionalen Emotionsraum aus. Thompson et al. (2007, S. 19) erheben, wie viel Zeit bis zur ersten Qualitätsbewertung einer Performance vom Beginn des Stücks an vergeht und berichten deutlich längere Verzögerungen von im Median 14s, welcher hier aufgrund von Ausreißern, die bedeutend länger brauchen, aussagekräftiger ist als der Mittelwert. Dabei sind die interindividuellen Unterschiede sehr groß mit einer Range von 12.26s und darüber hinaus variiert die Entscheidungsfindungszeit für die initiale Bewertung auch nach den Musikstücken und der Skala, also ob Gesamtqualität, Musikalität oder technisches Vermögen zu bewerten waren.

Welches Fazit sich bzgl. der Latenz für die Untersuchung von Performancebewertungen in einer tatsächlichen Konzertsituation ziehen lässt, ist schwierig zu beurteilen. Thompson et al. (2007) untersuchten zwar ebenfalls Performancebewertungen, allerdings nur anhand von Audio-Aufnahmen und sie messen die Entscheidungsfindungszeit zu Beginn der Stücke und treffen keine Aussage dazu, wie sich die Latenz im Zeitverlauf verhält, was vermutlich anhand der Daten auch nicht möglich wäre. Die übrigen Ergebnisse zur Latenz beziehen sich auf andere Inhalte, aber sind dafür nicht auf den Beginn eingeschränkt, sondern für beliebige Zeitpunkte anwendbar. Am genauesten wäre es, bei jedem Probanden einzeln die Reaktionszeit anhand einer ersten Bewertung oder Reaktionen auf markante Ereignisse zu bestimmen und damit weiter zu arbeiten, allerdings ist es methodisch auch nicht unproblematisch, Reaktionen bestimmten Ereignissen zuzuordnen, selbst wenn über die gewollten Bewertungen hinaus versehentliches Verrutschen u. Ä. ausgeschlossen werden könnte. Zudem könnte die Reaktionszeit zu bestimmten Zeitpunkten variieren, eine anfängliche Entscheidung möglicherweise länger dauern als kleinere Reaktionen im Zeitverlauf oder Zuhörer könnten während eines Experiments von ihrer Aufgabe abgelenkt werden, sich beispielsweise völlig in die Musik vertiefen, und daher stärker verzögert reagieren. Wenn möglich, sollte in dieser Situation eine anhand der Daten ermittelte oder geschätzte Latenzzeit gegenüber allgemeinen Angaben aus der Literatur vorgezogen werden.

2.6 Zusammenhang zwischen den kontinuierlichen und den retrospektiven Wertungen

Dieser Abschnitt beschäftigt sich damit, in welchem Verhältnis die abschließenden Gesamturteile zu den Echtzeit-Wertungen stehen. Dazu werden verschiedene Theorien sowie empirische Befunde vorgestellt, wobei der inhaltliche Fokus aufgrund fehlender Literatur nicht auf Performanceforschung beschränkt bleiben kann, sondern auch Theorien aus der (musikalischen) Emotionsforschung und der Medizin mit berücksichtigt werden und deren Übertragbarkeit diskutiert wird.

Die naheliegendste Annahme wäre, dass alle einzelnen Zeitpunkte gleich gewichtet in die Posthoc-Wertung einfließen, letztere also dem Mittelwert oder einer Summe der Einzelwerte entspricht. Der Unterschied zwischen diesen beiden Varianten wäre, dass bei der Summenbildung die Dauer der Erfahrung einen bedeutenden Einfluss hat, bei der Bildung eines Mittelwertes hingegen nicht. Eine längere gleichbleibend positive emotionale Erfahrung würde dann abschließend besser bewertet, als würde die gleiche Erfahrung bereits nach der Hälfte der Zeit enden. Bei der Mittelwerthypothese hingegen müsste in diesem Beispiel jeweils die gleiche Gesamtwertung herauskommen. Empirischen Befunden zufolge ist die Summenbildung ein schlechter Indikator zum Vorhersagen der Abschlusswertung und dieser Sachverhalt wird häufig als *Vernachlässigung der Dauer* (englisch: *duration neglect*) bezeichnet (vgl. Schubert, 2010, S. 244). Zur Mittelwertthese gibt es sowohl unterstützende als auch widerlegende Ergebnisse. So kamen beispielsweise Brittin und Sheldon (1995) zu dem Ergebnis, dass sie bei ihrer Erhebung von Präferenzurteilen in Echtzeit und posthoc nur für die Teilgruppe der Musikstudierenden zutraf, während die Nicht-Musikstudierenden retrospektiv im Schnitt einen Skalenpunkt niedriger bewerteten als durchschnittlich zuvor im Zeitverlauf. Brittin und Duke (1997) sowie Duke und Colprit (2001) kamen zu fast identischen Ergebnissen und jeweils zu dem Schluss, dass sowohl kontinuierliche als auch retrospektive Bewertungen von musikalischer Intensität bzw. Emotionsintensität für sich genommen sowohl inter- als auch intraindividuell eine hohe Konsistenz aufwiesen, allerdings die Abschlussbewertungen höher sind als die kontinuierlichen Ratings im Mittel. Dabei variierte der Abstand zwischen dem Mittelwert und dem „psychologischen Mittelwert“, wie Brittin und Duke (1997) die Posthoc-Bewertung bezeichnen, abhängig davon, wie stark die Emotionsintensität sich innerhalb eines der gehörten Musikstücke veränderte.

2.6.1 Peak-End-Rule

Infolge solcher Ergebnisse entwickelten sich Theorien, denen zufolge einzelne Zeitpunkte, wie der Anfang, das Ende oder der Moment mit der größten Intensität (meist *Peak* genannt), für die abschließende Bewertung, wichtiger als andere sind oder sogar alleinig ausschlaggebend. Redelmeier und Kahneman (1996) ließen ihre Probanden nacheinander erst eine Hand für 60 Sekunden in 14 °C kaltes Wasser halten, und dies dann mit der anderen Hand wiederholen, allerdings wurde dafür der Versuch um weitere 30 Sekunden verlängert und für diese Zeit die Wassertemperatur minimal auf 15 °C erhöht. Anschließend wurden sie gefragt, welches Experiment sie wiederholen wollten und die Mehrheit entschied sich für das zweite, welches länger ist, wobei also insgesamt mehr Schmerz ertragen werden muss. Ausschlaggebend war offensichtlich, dass das Ende etwas angenehmer war, was nahelegt, dass dieses einen größeren Einfluss auf die Erinnerung bzw. Bewertung der ganzen Episode hat. Während für dieses Experiment eine minimal-invasive Prozedur gewählt wurde, die nur für das Experiment durchgeführt wurde, untersuchten Kahneman, Fredrickson, Schreiber und Redelmeier (1993) das Erleben von unangenehmen medizinischen Behandlungen wie Darmspiegelungen oder Lithotripsien, der Zertrümmerung von Gallen- oder Nierensteinen. Hier wurde bereits ein Design mit der Kombination aus Echtzeitwertungen, die die Probanden während der Prozedur alle 60s abgaben, und einer abschließenden Bewertung angewandt. Bei beiden Erhebungen wurde dabei je die gleiche unipolare Skala für Schmerzempfinden („no discomfort“ bis „awful discomfort“) eingesetzt. Beim Vergleich der kontinuierlichen Bewertungen mit den Schlussbewertungen fiel auf, dass der schlimmste Moment und das Ende der Prozedur, wofür die letzten drei Minuten berücksichtigt wurden, am stärksten mit dem erinnerten Schmerzempfinden korrelierten. Ein Modell, dass neben diesen beiden Größen, auch noch den Anfang, den Durchschnitt, die Summe der Bewertungen sowie die Dauer der Prozedur miteinbezog, erklärte die Bewertungen nur minimal besser, z. T. nicht signifikant, als das einfachere Modell und die Dauer der Prozedur korrelierte nicht mit der globalen Schmerzbewertung. Kahneman et al. (1993) schließen daraus, dass einerseits eine Vernachlässigung der Dauer vorliegt und das erinnerte Schmerzempfinden vorwiegend den schlimmsten Moment sowie das Ende reflektiert. Dieser Zusammenhang, dass vorwiegend die Bewertung des Peak-Moments und des Endes eines Erlebnisses ausschlaggebend für die retrospektive Gesamteinschätzung sind, wird als *Peak-End-Theorie* bezeichnet, welche nicht ausschließlich in der Medizin angewandt bzw. überprüft wird.

Fredrickson und Kahneman (1993) zeigten ihren Versuchspersonen aversive und angenehme Filmszenen und ließen die Probanden in einem ersten Experiment kontinuierlich sowie retrospektiv bewerten, wie (un-)angenehm sie die Szenen fanden, während in einem zweiten Experiment die Teilnehmer erst nach dem Sehen der Szenen gebeten wurden einzuschätzen, wie stark diese zu einer (un-)angenehmen Gesamterfahrung beitragen. Die Ergebnisse unterstützen die Peak-End-Theorie und nur für negative Szenen wurde ein kleiner Zusammenhang mit der Dauer gefunden, und dieser konnte komplett dadurch erklärt werden, dass sich der negative Affekt im Verlauf der Szene steigerte. Die Studie ist interessant, weil das Design den Vergleich zwischen Gesamteinschätzungen mit und ohne vorhergehende kontinuierliche Evaluation zulässt, allerdings wurden in den Experimenten verschiedene Skalen benutzt, sodass ein direkter Vergleich nicht möglich ist. Retrospektive Wertungen aus Studie 1 und Studie 2 korrelieren mit $r > .8$, für die aversiven Filmszenen sowohl für die Bedingung in Studie 2 mit genauer Versuchsanweisung vor dem Sehen der Videoclips, als auch wenn diese erst anschließend gegeben wurde. Für die positiven Videoclips hingegen traf dies nur zu, wenn die Zuhörer vorab informiert wurden, und bei der Versuchsbedingung mit verzögerter Information über die Bewertungsaufgabe war die Korrelation mit den Ergebnissen aus Studie 1 sowie denen unter der anderen Bedingung in Versuch 2 deutlich niedriger ($r = .48$ bzw. $r = .58$). In diesem Zusammenhang ließ sich also zeigen, dass die kontinuierliche Bewertung keinen starken verzerrenden Einfluss auf die Abschlussbewertung hatte, sofern die Probanden bereits vorher wussten, worauf sie beim Zusehen achten sollten.

Die Untersuchung der retrospektiven Evaluation schmerzhafter Prozeduren hat zunächst nichts mit ästhetischem Erleben oder dem Bewerten von musikalischen Leistungen zu tun, aber das Erleben von Filmszenen geht bereits in die richtige Richtung. Einen Versuch, die Peak-End-Theorie für Musik als Stimulus zu überprüfen, unternahmen Rozin, Rozin und Goldberg (2004) indem sie in einem ähnlichen Design wie Kahneman et al. (1993) und Fredrickson und Kahneman (1993) in der ersten Studie untersuchten, wie ihre Probanden zu einer abschließenden Bewertung der Emotionsintensität für eine Auswahl von Musikstücken kommen. Ihre Ergebnisse unterstützen grundsätzlich die Peak-End-These, geben allerdings als dritten wichtigen Einflussfaktor Momente an, deren Emotionsintensität bedeutend höher ist, als die der direkt vorhergehenden Momente. Dabei ist allerdings der Peak-Moment der wichtigste. Aufgrund des halben Datensatzes entwickelten Rozin et al. (2004) ein Modell, dass eine gewichtete Summe aus allen Zeitpunkten bildet, und zwar so dass eben Peaks, das Ende und eine hohe Steigung ein stärkeres Gewicht

bekommen, als die übrigen. So wird durchaus die Dauer zu einem gewissen Teil mit berücksichtigt. Das Modell wird auf die zweite Hälfte der Daten angewandt um es zu überprüfen und gemessen mit t-Tests ist es ein signifikant besserer Prädiktor als Ende, Mittelwert, Peak oder Peak-End. Eine Regressionsanalyse hingegen zeigt, dass das Modell nur gleichwertig ist zu einer Kombination von Summe, Peak und Ende. Rozin et al. (2004) argumentieren, dass das Modell theoretisch dennoch zufriedenstellender sei, da sie es für unwahrscheinlich halten, dass der Hörer tatsächlich in der Lage sei, jeden einzelnen Moment exakt in Erinnerung zu behalten und am Schluss zu evaluieren.

Schäfer et al. (2014) untersuchen ebenfalls die Übertragbarkeit des Peak-End-Modells bzw. eines Modells, das mehrere Peaks und das Ende berücksichtigt, auf momentan wahrgenommene Emotionsintensität bzw. die retrospektive Einschätzung der Emotionsintensität von Musikstücken und argumentieren durchaus kritisch, dass die Mehrzahl der bisherigen empirischen Befunde, die die Peak-End-Theorie unterstützen, sich auf negative Erfahrungen beziehen. Speziell in Frage gestellt wird, inwiefern die Verarbeitung von schmerzhaften Erfahrungen mit dem Musikerleben vergleichbar sei. Das Design der hier präsentierten Studie entspricht dem von Rozin et al. (2004), wobei wieder die Emotionsintensität für musikalische Stimuli gemessen wird. Für die Datenanalyse wurde der globale Peak als die höchste Bewertung im gesamten Zeitverlauf definiert, weitere lokale Maxima als Wertungen, die in einem Fenster von ± 20 s die höchsten waren und darüber hinaus oberhalb des gleitenden Mittelwerts für ein 20 s-Fenster lagen, und die End-Wertung wurde berechnet als Mittelwert aller Wertungen der letzten 10 s. Zum Überprüfen der Peak-End-Theorie wurde der Mittelwert aus dem globalen Peak und dem End-Wert berechnet, bzw. als Alternative der Mittelwert aller lokalen Peaks und des End-Wertes. Es wurden noch diverse weitere mögliche Prädikatoren einbezogen, die im Vergleich zu den genannten für das Ergebnis weniger wichtig waren, wovon einerseits noch die Variation genannt sei, die als Standardabweichung über das Zeitprofil definiert wurde, da diese in die hier zitierten, weiteren Analysen miteinbezogen wurde. Andererseits erwähnenswert für die Datenanalyse in dieser Arbeit ist die Anfangswertung, die als arithmetisches Mittel der Wertungen des Zeitraums von 5 bis 15 s nach Beginn des Stücks berechnet wurde, da 5 s Latenz für emotionale Reaktion und Abgabe einer entsprechenden Selbstauskunft vermutet werden (vgl. auch Nagel, 2007). Die Korrelationen zwischen den drei Modellen Mittelwert, Peak-End und mehrere Peaks-End sind alle hoch, knapp am höchsten für den Mittelwert, und um das Ergebnis nicht lediglich auf kleine Unterschie-

de des Korellationskoeffizienten zu stützen, werden multiple Regressionsanalysen durchgeführt. Aufgrund von Interkorellationen zwischen verschiedenen möglichen Prädiktoren konnten nicht alle Parameter gleichzeitig in eine solche Analyse miteinbezogen werden. Insofern wurde die Priorität auf die zu prüfende Theorie gesetzt und je eine Analyse mit Peak-End bzw. eine mit mehreren Peaks-End und außerdem dem Mittelwert, der Variation über die Zeit und die Anzahl der Peaks berechnet. Dabei ergab sich bei einer Varianzaufklärung von je ca. 70% bei beiden Modellen der größte Regressionskoeffizient für den Mittelwert ($\beta = .59$ bzw. $\beta = .58$) und ein kleinerer Koeffizient für Peak-End ($\beta = .26$) bzw. mehrere Peaks-End ($\beta = .25$), während Varianz und Anzahl der Peaks demnach keine signifikanten Prädiktoren für die Gesamteinschätzung waren. Schäfer et al. (2014) folgern, dass eine ausschließliche Betrachtung von Peak- und Endmoment nicht ausreichend sei, wenngleich diese beiden Momente offenbar einen besonders starken Einfluss hätten. Ein Modell, das mehrere Peaks berücksichtige, biete keinen Mehrwert gegenüber dem simplen Ein-Peak-End-Modell und grundsätzlich sei das Modell theoretisch für die Anwendung auf Musik noch nicht zufriedenstellend. Beispielsweise sei noch unklar, in welchem Verhältnis Peak-Wertungen und Chills stehen und wie die Ergebnisse zu interpretieren seien, ob beispielsweise die Hörer erst nachträglich dem Peak und Ende mehr Ausschlagkraft zuweisen, oder dies bereits während des Hörens passiert. Die Argumentation, ein *slope effekt*, den Rozin et al. (2004) fanden und demnach starke Wertungsanstiege zusätzlich großen Einfluss auf die Posthoc-Wertung haben, liege hier nicht vor, da die Variation zwar eine mittlere Korrelation mit den retrospektiven Bewertungen habe, aber bei den Regressionsanalysen keinen Einfluss habe, ist zu einfach, da nicht tatsächlich das Gleiche gemessen wird.

Insgesamt ist die Anwendbarkeit der Peak-End-Theorie auf Musikwahrnehmung weiterhin nicht gesichert, speziell die beiden musikbezogenen Studien kommen zu dem Ergebnis, dass Erweiterungen des Modells, die weitere Parameter mit einbeziehen, oder doch der Mittelwert über den gesamten Verlauf überlegen sind. Die Theorie, welche Prozesse überhaupt dazu führen, dass bestimmte Momente stärkeren Einfluss auf die erinnerte Emotionsintensität haben, ist umstritten. Es stellt sich die Frage, ob eine retrospektive Bewertung in jedem Fall mit der Erinnerung gleichzusetzen ist, oder ob nachträgliche Evaluationsprozesse das Bild verzerren. Speziell für die Studie, die im Rahmen dieser Arbeit durchgeführt wurde, müsste die Theorie von einer Selbstauskunft über die Emotionsintensität weiter auf den Prozess einer Performancebewertung übertragen werden. Beim Bilden einer Gesamtnote ist zu vermuten, dass die Zuhörer eher noch weniger versuchen, alle Zeitpunkte zu einer

Bewertung zusammenzufassen, sondern viel bewusster eine Gewichtung vornehmen, indem sie beispielsweise einen Aussetzer, den sie zu dem Zeitpunkt zwar sehr schlecht bewertet haben, komplett unberücksichtigt lassen, weil sie der Rest des Stücks überzeugt hat. Man könnte in diesem Fall auch die Bildung von „pädagogischen Noten“ erwarten, die Bewertung stark als Feedback für den Performer sehen und somit keine sehr schlechte Note geben um nicht zu demotivieren oder noch nicht in einem informelleren Kontext vor einer Abschlussprüfung die Bestnote ziehen, um weiterhin einen Ansporn zu schaffen bzw. zu verhindern, dass jemand enttäuscht ist, wenn es dann, wenn es darauf ankommt, im Vergleich schlechter bewertet wird.

Ebenfalls problematisch für die Übertragung ist, dass in allen Studien sowohl im medizinischen Bereich als auch im Bezug auf emotionales Erleben von Filmclips oder Musik eine unipolare Skala verwendet wurde, um die Schmerzintensität oder Emotionsintensität zu messen, nicht aber verschiedene Emotionsqualitäten. Speziell bei Emotionen könnte es ausschlaggebend sein, zwischen negativ und positiv zu unterscheiden, zumal Fredrickson und Kahneman (1993) für aversive bzw. positive Filmclips zu unterschiedlichen Ergebnissen kommen, aber das Problem wurde bisher nicht gelöst, sondern nur durch eine reine Intensitätsskala umgangen. Bei einer Bewertung einer Performance ist es aber denkbar, dass sowohl besonders brillante oder emotional präsentierte Stellen besonders wichtig sind, aber auch, dass Momente mit schlechter Intonation, fehlendem Spannungsbogen oder ein Aussetzer sehr gut in Erinnerung bleiben bzw. aus anderem Grund stärker gewichtet werden bei der Evaluation. Diese beiden Möglichkeiten existieren unabhängig von der Skala, insofern wird es schwierig sein, das Problem in diesem Kontext auszuschließen und eine Lösung muss gefunden werden.

2.6.2 Primacy Effekt

Eine andere Theorie, die davon ausgeht, dass ein bestimmter Zeitpunkt am ausschlaggebendsten ist, ist die Annahme, dass der erste Eindruck zählt, also ein sogenannter Primacy Effekt vorliegt. Auch im Alltagsverständnis wird oft die Annahme bzgl. der Bewertung praktischer Prüfungen in Musik, beispielsweise Aufnahmeprüfungen, getroffen, derzufolge im Wesentlichen die ersten wenigen Takte zählen und dann das Ergebnis bereits mehr oder weniger feststehe. In der Literatur ist diese Theorie hingegen schlecht belegt und besonders musikbezogen finden sich dazu kaum Studien.

Platz und Kopiez (2013) deuten gewissermaßen an, die Bewertung des Bühnenauftretens habe einen Einfluss auf die Performance-Wahrnehmung insofern,

dass hier die Voraussetzungen geschaffen werden, dass der Zuhörer überhaupt motiviert wird, aktiv zuzuhören. Eine abschließende Bewertung wurde allerdings nicht erhoben, sodass nicht belegt werden kann, ob die angegebene Motivation zum Weiterhören bzw. auch direkt die Einschätzung der Angemessenheit des Bühnenauftritts in einem Zusammenhang mit der Bewertung des gesamten Auftritts steht. Außerdem gibt es nur vereinzelt Fälle, bei denen der Bühnenauftritt tatsächlich als unangemessen bewertet wurde, sodass die Probanden angaben, nicht weiter hören zu wollen.

Wapnick et al. (2009) untersuchten den Einfluss visueller Attribute wie Attraktivität und Kleidung auf die Performancebewertung in drei Versuchsbedingungen, in denen die Probanden jeweils die ersten 25, 55 oder 115 s einer auf Video aufgezeichneten musikalischen Darbietung gezeigt bekamen. Die Ergebnisse unterschieden sich nach Geschlechtern, sodass bei Musikerinnen eine als höher eingeschätzte Attraktivität einen positiven Einfluss auf die Bewertung ihrer Performance hatte, bei männlichen Musikern hingegen die Kleidung ausschlaggebend war, allerdings jeweils nur bei der kürzesten Hördauer von 25 s. Bei den längeren Ausschnitten spielten diese visuellen Attribute keine Rolle, insofern lässt sich hier ableiten, dass ein positiver visueller erster Eindruck nicht über die Zeit erhalten bleibt, sondern lediglich dann einen Einfluss auf die Performancebewertung hat, solange nur wenige musikalische Informationen als Basis für die Bewertung zur Verfügung stehen. Auch wenn eine Einschränkung auf den Einfluss guten Aussehens stattgefunden hat, widersprechen die Ergebnisse eher der These, dass vorwiegend der Beginn einer Performance ausschlaggebend ist, da offensichtlich der weitere Verlauf des Stückes stark berücksichtigt wurde.

Möglicherweise lassen sich auch allgemeinere Erkenntnisse zum ersten Eindruck von einer Person auf einen Live-Auftritt eines Musikers übertragen, zumal nach Platz (2014) eine erste Evaluation bereits vor dem ersten klingenden Ton erfolgen kann. Ybarra (2001) untersuchte, inwiefern ein erster Eindruck von einer Person bestehen bleibt oder nachträglich angepasst wird, und kam dabei zu dem Ergebnis, dass es einen Unterschied macht, ob der erste Eindruck positiv oder negativ war. Bei einem positiven Eindruck bleiben die Personen eher offen, lassen sich also auf weitere Eindrücke bzw. Informationen ein und integrieren diese in ihr Bild von einem Gegenüber. Ist der erste Eindruck hingegen negativ, ist es unwahrscheinlicher, dass weitere Eindrücke offen berücksichtigt werden, und stattdessen bleibt eher das erste Bild von einer Person isoliert und unverändert stehen. Es ist durchaus denkbar, dass diese Annahme auch für musikalische Darbietungen relevant ist und

das Fehlen musikbezogener Ergebnisse darauf zurückzuführen ist, dass bislang vorwiegend Emotionen, aber kaum Bewertungen der Darbietungsqualität mit *continuous response*-Verfahren untersucht wurden, sich diese beiden Phänomene aber unterschiedlich verhalten im Bezug auf die Erinnerung bzw. Meinung, die nach dem Hören bleibt. Die Implikation wäre in diesem Fall, dass wer als Musiker seinen Auftritt schlecht beginnt, dies später nicht wieder aufholen kann, hingegen wer einen guten Anfang schafft, weiterhin eine hohe Leistung bringen muss, um tatsächlich gut bewertet zu werden.

2.6.3 Evolution der Bewertung

Eine völlig andere These zur Interpretation des Zusammenhangs zwischen Echtzeit-Wertungen und Gesamturteil vertreten Thompson et al. (2007), die sich in ihrer Studie auch konkret mit Performancebewertung beschäftigen. Sie bezeichnen ihre Theorie als *Evolution der Bewertung* (engl. „evolution of judgement“) und verstehen darunter folgenden Meinungsbildungsprozess:

„As soon as some minimum amount of musical information has been heard, an initial decision is made. This has the status of a ‚working hypothesis‘ about the quality of the performance, which is subject to modification and adjustment as the performance continues. At some time before the end of the piece a stable point is reached, reflecting ‚finalization‘ of the judgment; this is then translated into the summary evaluation (Thompson et al., 2007, 13f).“

Demnach liegt bei der abgegebenen kontinuierlichen Bewertung keine momentane Einschätzung vor, sondern stattdessen eine Evaluation des gehörten Stücks bis zu dem aktuellen Zeitpunkt, sodass folglich die letzte Echtzeit-Bewertung bereits inhaltlich der Abschlusswertung entspricht. Dafür, wie viel Zeit bis zur ersten Bewertung vergeht, liegt keine Hypothese vor, allerdings für die Beendigung der Wertung erwarten Thompson et al. (2007), dass diese spätestens nach 90 s stattfindet.

Thompson et al. (2007) gehen in ihrer eigenen Studie von drei möglichen Zusammenhängen zwischen den kontinuierlichen Bewertungen und den Gesamt-Ratings aus, nämlich dass letzteres dem Mittelwert der Echtzeit-Wertungen entspricht, ein Recency-Effekt vorliegt, also das Ende des Stückes besser in Erinnerung bleibt, oder eben die eigene Theorie der Evolution der Bewertung. Dabei untersuchen sie auch die Entscheidungsbildungszeit für die erste Bewertung, sowie die Häufigkeit

und das Ausmaß von Bewertungsänderungen und vergleichen die mittleren Verläufe der kontinuierlichen Bewertungen für drei verschiedene verwendete Skalen (vgl. Abschnitt 2.4.3). 33 Probanden, die alle mit klassischer Musik zu tun hatten dadurch, dass sie überwiegend Musikstudierende waren oder ansonsten regelmäßige Konzertgänger, bewerteten Aufnahmen mehrerer Varianten von zwei Klavierstücken in Echtzeit und jeweils noch einmal global nach jedem Stimulus. Dabei wurde für die Abschlussbewertung eine 7-stufige Skala verwendet, die auch für die Echtzeit-Wertungen an der dafür verwendeten kontinuierlichen Skala angezeigt wurden. Abschließend bewerteten die Versuchsteilnehmer jeweils die Skalen Gesamtqualität, technisches Können und Sicherheit, sowie Musikalität. Kontinuierlich hingegen wurden sie in drei Gruppen aufgeteilt und bewerteten pro Gruppe immer eine der Skalen.

Die Entscheidungsfindungszeit bis zur ersten Wertung betrug für die Versuchsbedingung Gesamtqualität im Mittel 18.09 s, wobei trotz Bereinigung der Ausreißer der Median mit 15.00 s etwas niedriger ist und der Wertebereich nur als Range von 12.16 s angegeben ist, sodass die tatsächlichen Maxima und Minima unbekannt sind. Die Ergebnisse für diese Bedingung variierten ein wenig für die beiden verschiedenen Stücke und für die anderen Skalen ergaben sich vergleichbare Entscheidungsfindungszeiten. Bezüglich der Bewegungsintensität kamen Thompson et al. (2007) zu dem Ergebnis, dass die Anzahl der Bewegungen pro Minute über den Zeitraum zunächst ansteigt und erst kurz vor Ende der Stücke wieder abnimmt, wobei graduelle Veränderungen der Bewertung nicht mitgezählt wurden und eine Wertung erst ab einer Dauer von 2 s als „stabil“ angesehen wurde. Das mittlere Ausmaß pro Bewegung sinkt hingegen bereits nach 60 s (Gesamtqualität und Musikalität) bzw. 90 s (Technische Souveränität) ab und der Verlauf der Kurve der mittleren Bewertungen (außer für technisches Können), die vorher steigend waren, flacht ab, was die Autoren als Beendigung der Wertung interpretieren und damit eine Teilhypothese für Evolution der Bewertung bestätigt sehen.

Mit einer Anova mit Messwiederholung für die drei Bewertungen für Anfang und Ende des Stücks, sowie Gesamtbewertung wurde ein signifikanter globaler Effekt gefunden und die Berechnung von wiederholten Kontrasten ergab eine positive Differenz zwischen Anfangs- und Endwertung, hingegen aber keinen signifikanten Unterschied zwischen End- und Gesamtwertung. Im gesamten Artikel werden weder Effektgrößen berechnet noch Power-Analysen durchgeführt, wenn Testergebnisse nicht signifikant sind, und diese fehlenden Größen können aufgrund der angegebenen Werte auch nicht nachträglich berechnet werden. Dennoch ist festzuhalten, dass die

Gruppen bei 33 Probanden, die auf drei Bedingungen aufgeteilt wurden, innerhalb einer Gruppe selbst für eine ANOVA mit Messwiederholung wenige Probanden sind, insofern die Nicht-Signifikanz nicht unbedingt damit gleichgesetzt werden kann, dass kein Effekt vorliegt. Thompson et al. (2007) argumentieren, dass aufgrund des Effekts zwischen Anfangs- und Endwertung ein Mittelwertzusammenhang zwischen kontinuierlichen und abschließenden Wertungen unwahrscheinlich sei, was selbst, wenn man den später im Fazit erwähnten, wenn auch nicht in Zahlen belegten, Unterschied zwischen Anfangs- und Abschlusswertung miteinbezieht, etwas dürftig ist. Selbst wenn die Anfangswertung zunächst deutlich niedriger war, kann im Mittel noch die Endwertung erreicht werden. Erst unter Berücksichtigung der Verläufe der Wertungen über die Zeit, die praktisch durchgängig steigend sind, kann es einigermaßen ausgeschlossen werden. Dieses Argument fehlt aber in diesem Kontext bzw. wird nur auf die Beendigung der Wertung und somit die Abwägung zwischen einem Recency-Effekt und einer evolvierenden Bewertung bezogen. Eine eindeutigere Widerlegung wäre hier angebracht gewesen, da individuelle Bewertungsverläufe auch von dem gemittelten Bild abweichen können. Dennoch ist in diesem Fall ein Recency-Effekt bzw. die Evolution der Bewertung wahrscheinlicher und die Autoren erläutern, dass hier beide Theorien zu den Ergebnissen passen und daher nicht entschieden werden kann, welche Interpretation dem Meinungsbildungsprozess besser entspricht. Sie tendieren zum Modell der Evolution der Bewertung, da die Wertung sich spätestens nach 90 s praktisch nicht mehr verändert, scheinbar abgeschlossen ist.

Diese Unterscheidung ist ein generelles Problem, da es sich nicht zwingend um ein mit Daten messbares Phänomen handelt, sondern eher um eine Uminterpretation. Wie soll man belegen, ob die Zuhörer – willentlich oder unbewusst – statt eine momentane Bewertung abzugeben, bereits das vorher gehörte integrieren? Die Argumentation mit der Beendigung der Bewertung kann zwar ein Indiz sein und ist bei einem reinen Recency-Effekt auch unwahrscheinlicher. Unter valideren Bedingungen in einer Live-Performance-Situation oder auch bereits mit Audio- oder Videoaufnahmen von tatsächlichen Live-Performances, bei denen nicht nachträglich die Fehler verbessert wurden, könnten die Ergebnisse bereits anders aussehen. Beispielsweise könnte die Zeitspanne, bis eine Bewertung abgeschlossen ist, deutlich länger sein oder die Bewertung wird nicht vor Ende des Stücks abgeschlossen. Dennoch haben die Zuhörer möglicherweise die gesamte bisherige Performance statt nur die letzten wenigen Sekunden im Blick. Dies ist auch eine Frage der Auslegung des Begriffs *Evolution*: Muss sie abgeschlossen werden, oder kann sie ein

Prozess sein, der immer weiter läuft und dann ggf. vom Ende des Stimulus beendet bzw. unterbrochen wird? Eventuell sollte insofern der Begriff einer akkumulierten Bewertung, wie Platz (2014) ihn für dieses Phänomen verwendet, bevorzugt werden.

Es bleibt festzuhalten, dass die Theorie der Evolution der Bewertung von dieser ersten Studie von Thompson et al. (2007), die konkret kontinuierliche Performancebewertung untersucht, unterstützt wird und eine durchaus plausible Erklärung für die Ergebnisse ist, insofern in diesem Rahmen weiterhin überprüft und diskutiert werden sollte.

3 Ziele und Hypothesen

Das Ziel dieser Arbeit, die Konsistenz der Performance-Ratings durch das Publikum zu untersuchen, wird anhand der drei, bereits in der Einleitung aufgeführten Teilfragen bzw. -ziele verfolgt. Aus der vorgestellten Literatur ergeben sich in vielen Fällen konkrete Hypothesen bzgl. der Ergebnisse, die an dieser Stelle zusammengefasst werden. Darüber hinaus lassen sich teilweise noch Hypothesen bilden, die nicht direkt von der Literatur unterfüttert werden, aber sich aus der Erfahrung am Institut für Musikwissenschaft und Musikpädagogik bzw. aus dem Alltagsverständnis heraus gut begründen lassen, sodass sie mit aufgenommen werden. Allerdings hat die Methode der kontinuierlichen Erhebung von Bewertungen einer musikalischen Darbietung insbesondere im Kontext von Live-Performances noch explorativen Charakter, sodass nicht überall vorab begründete Thesen gebildet werden können und stattdessen offen an die Analyse herangegangen wird. Dafür werden statt Hypothesen spezifische Unterfragen gebildet, die genau angeben, was im Hinblick worauf betrachtet werden soll, allerdings keine Annahme über das Ergebnis treffen.

1. Wie konsistent sind die retrospektiven Gesamturteile?

Dieser Absatz listet die Hypothesen, die sich ausschließlich auf die Analyse der Abschlussbewertungen nach dem Hören des jeweiligen Stückes.

- a) Die Dozenten bewerten interindividuell reliabel.
- b) Das gesamte Publikum bewertet interindividuell reliabel.
- c) Mit zunehmender Expertise im Bewerten, sind die Zuhörer kritischer, geben also tendenziell schlechtere Bewertungen ab.
- d) Die Selbsteinschätzungen der Musiker sind nicht konsistent mit den Publikumsbewertungen.

Die Hypothesen, die sich auf das Bewertungsverhalten der Zuhörer beziehen, sind positiv formuliert, es wird also grundsätzlich angenommen, dass eine hohe Übereinstimmung zwischen den Bewertungen der Probanden besteht, eine Performance also (einigermaßen) einheitlich bzw. objektiv bewertet werden kann. Die in Abschnitt 2.4.2 zitierten Studien belegen entsprechende

Übereinstimmungen, meist in Form von Korrelationen, die mittel bis hoch sind, unabhängig davon ob die Juroren besonders erfahren waren, oder die Peers der Musiker die Bewertungen vornahmen. Die Hypothese 1c lässt sich nicht aus der Literatur heraus begründen, es ist aber naheliegend, dass beispielsweise Dozenten, die häufig Performances bewerten müssen, einen höheren Anspruch haben könnten, als Studienanfänger. Die letzte Hypothese in diesem Absatz stützt sich auf die Ergebnisse von Bergee (1993) die ebenfalls in Abschnitt 2.4.2 bereits dargestellt wurden.

2. Wie konsistent sind die kontinuierlichen Bewertungen?

Dieser Absatz fasst Ziele der Analyse der kontinuierlichen Daten im Bezug auf Konsistenz zusammen. Können anhand der kontinuierlichen Bewertungen Rückschlüsse auf die Performance gezogen werden, um beispielsweise festzustellen, auf welche Fehler das Publikum reagiert und auf welche nicht? Mit diesem Ziel, Verbindungen zur Aufzeichnung der Performance herzustellen, werden die kontinuierlichen Bewertungen auf Konsistenz hin untersucht. Ein Abgleich mit jeder einzelnen Bewertung wäre nicht praktikabel, von einer mittleren Bewertung auszugehen allerdings nur dann zielführend, wenn im Publikum ein gewisser Konsens bzgl. der Güte der Performance zu bestimmten Zeitpunkten oder deren Entwicklung im Zeitverlauf besteht.

- a) Wie stark schwankt die Standardabweichung?
- b) Sind die kontinuierlichen Bewertungen interindividuell reliabel?
- c) Wann im Verlauf der Performances gibt es Zeitphasen mit besonders hoher oder niedriger Einigkeit?
- d) Gibt es unabhängig von der absoluten Bewertung einen gleichförmigen Verlauf?
- e) In welchen Zusammenhang mit den Aufnahmen lassen sich die Verläufe sowie Phasen niedriger und hoher Übereinstimmung bringen?

Da Performancebewertungen bislang kaum in Echtzeit-Verfahren untersucht wurden, jedenfalls nicht im Hinblick auf die Konsistenz der Bewertungsverläufe, lassen sich dafür keine Hypothesen ableiten. Ergebnisse aus der Forschung zum emotionalen Erleben von Musik zu übertragen wäre an dieser Stelle fragwürdig. Daher findet diese Analyse nicht hypothesenbasiert statt, sondern ist eher als explorative Sichtung der Daten zu verstehen.

3. Wie hängen kontinuierliche und retrospektive Wertung zusammen?

Diese Forschungsfrage lässt sich basierend auf der Literaturrecherche präzisieren zu: Welche der in Abschnitt 2.6 vorgestellten Modelle sind in der Lage den Zusammenhang zwischen den Echtzeit-Bewertungen und den Gesamturteilen zu erklären?

- a) Mittelwert: Der Mittelwert der kontinuierlichen Bewertung erklärt die Gesamtbewertung am besten.
- b) Peak-End-Rule: Der Mittelwert aus Peak-Wert und End-Wert der kontinuierlichen Bewertung erklärt die Gesamtbewertung am besten.
- c) Primacy Effekt: Die Bewertung des Anfangs der Performance erklärt die Gesamtbewertung am besten.
- d) Evolution der Bewertung: Es liegt kein tatsächliches momentanes Urteil vor, sondern stattdessen eine Beurteilung der Performance bis zu dem entsprechenden Zeitpunkt.
 - (i) Die Bewertung des Endes des Stückes entspricht der Gesamtbewertung am besten.
 - (ii) Nach ca. 90 Sekunden ist das endgültige Urteil gefunden und verändert sich kaum mehr.
- e) Die Bewertung des Bühnenauftrittsverhaltens hat Einfluss auf die Bewertungen zu Beginn eines Stückes, aber nicht darüber hinaus.

Es wurden der Mittelwert, die Peak-End-Theorie und die Theorie der Evolution der Bewertung aufgenommen, da diese sich schon in anderen Studien bestätigt haben, und zusätzlich der Primacy Effekt, obwohl es dafür bislang keine empirischen Belege gibt, dieser aber in den Alltagstheorien sehr präsent ist. Mehrere Modelle könnten einen Zusammenhang zwischen den kontinuierlichen und retrospektiven Bewertungen erklären. Welches Modell erklärt aber den Zusammenhang am besten und ist insofern gegenüber den anderen Modellen als überlegen anzusehen? Um diese Frage zu beantworten, sind alle Hypothesen so formuliert, dass die Modelle als „das Beste“ getestet werden. Für die Evolution der Bewertung wird zusätzlich eine weitere Hypothese zur Beendigung der Bewertung gebraucht, deren Ablehnung allerdings, sofern die erste Hypothese bestätigt werden kann, nicht grundsätzlich zu einer Ablehnung des Modells führen würde, sondern lediglich eine Umdefinition

erforderlich machen. Außerdem soll geprüft werden, ob das Bühnenauftrittsverhalten sich stark auf die Performancebewertung auswirkt, um an die Ergebnisse von Wapnick et al. (2009), und Platz (2014) anzuschließen, dass einerseits der erste Eindruck wichtig sei, andererseits aber visuelle Einflüsse bei längerer Hördauer unwichtig werden.

4 Methode

4.1 Rahmen des Experiments

Die Studie wurde im Januar 2015 im Rahmen eines institutsinternen Studierendenvorspiels durchgeführt, in dem ein Teil der teilnehmenden Musiker von einem Teil des Publikums in Echtzeit sowie nach jedem Stück abschließend bewertet wurde. An diesen Vorspielen müssen alle Studierenden, sofern sie Instrumentalunterricht erhalten, mit Ausnahme der Erstsemester, innerhalb des Semesters einmal mit ihrem Hauptinstrument teilnehmen. Die Vorspiele sind unbewertet insofern können den Teilnehmern durch die Studie keine für das Studium relevanten Nachteile entstehen und die Wahl der Stücke ist weitgehend frei, auch wenn es erwünscht ist, dass Stücke gespielt werden, die auch für eine Instrumentalprüfung geeignet wären. Die Moderation der Vorspielabende übernimmt je eine Gruppe Studierender im ersten Semester.

4.2 Stichprobe

4.2.1 Musiker

Insgesamt nahmen $N_M = 12$ Musiker am Experiment teil, die solistisch oder als Ensemble, bis zu Quartettgröße, fünf Musikstücke darboten (Details s. Tabelle 4.1). Alle für das Vorspiel angemeldeten Musiker wurden persönlich angeschrieben, ob sie bereit seien, an der Studie mit dem Stück, das sie sowieso dort spielen wollten, teilzunehmen und es erklärten sich insgesamt gut die Hälfte dazu bereit. Es sollten maximal vier bis fünf Stücke vom Publikum bewertet werden, um den Aufwand für die Bewertenden vertretbar zu halten und, da aufgrund des explorativen Charakters der Studie eine Auswertung noch weiterer Stücke gegenüber der Einschränkung auf einige wenige nicht gewinnbringend erschien. Die Auswahl der Stücke wurde alleine aufgrund der Spieldauer, nicht aufgrund der erwarteten Performance o. Ä., getroffen, sodass zwei Pianistinnen, die ein neun- bzw. fünfzehn-minütiges Stück spielen wollten, abgesagt wurde. Somit blieben die in Tabelle 4.1 aufgeführten Stücke übrig, die jeweils nicht länger als fünf Minuten dauerten. Das Vorspiel begann

Nr.	Komponist	Titel	Instrumente
1	Carl Reinecke	Konzert für Flöte und Orchester D-Dur, II. Lento e mesto	Flöte, Klavier
2	Giovanni Battista Pergolesi	Stabat Mater	2 x Gesang, Klavier
3	Georg Philipp Telemann	Duett Nr. 3 – I. Vivace & II. Poco Presto	2 Altblockflöten
4	Sergej Rachmaninov	Etudes Tableaux, op. 33 – No. 8 in g-Moll	Klavier
5	Tommy Emmanuel	Tall Fiddler	Gitarre, Geige, Bass, Akkordeon

Tabelle 4.1: Programm am Vorspielabend

mit diesen fünf Beiträgen und danach spielten weitere Musiker bzw. Ensembles, die nicht bewertet wurden. Alle Musiker waren Studierende des Instituts, die auf ihrem Hauptinstrument musizierten, ihr Instrument also schon viele Jahre spielen mit Ausnahme des Bassisten, der in diesem Semester mit Kontrabass angefangen hatte und eigentlich E-Bass spielt. Die Studierenden waren zwischen 20 und 24 Jahre alt ($M = 22.25$) und neun von ihnen waren weiblich, was durchaus der Geschlechterverteilung der Studierenden im Institut für Musikwissenschaft und Musikpädagogik entspricht. Sieben waren im Bachelor eingeschrieben (ab dem dritten Semester) und fünf im Master, je mit dem Ziel Lehramt.

4.2.2 Publikum

Im Publikum wurden von 28 Personen mit Hilfe eines iPads Bewertungen vorgenommen. Ausgewertet werden nur die Daten von $N_P = 27$ Personen, da ein Teilnehmer sich nicht als Musikstudierender oder Dozent eingestuft, sondern in der Rubrik *Sonstiges* „Maurer“ angegeben hat und daher unklar ist, ob er als externer hinzu kommt oder durchaus Musikstudierender ist, nur die Frage mehr auf seine bisherige Bildungskarriere bezogen hat. Von den 27 Teilnehmern waren 5 Dozenten, 19 Studierende (13 im Bachelor und 6 im Master) sowie 3 Absolventen eines Master of Education, die nicht mehr an der Universität sind, im Folgenden bei Gruppen aber immer mit zu den Masterstudierenden gezählt werden, da ihr Abschluss noch nicht lange zurück liegt (Details s. Abb. 4.1). Durch diese Stichprobenszusammensetzung ergibt sich eine Altersspanne von 20 bis 53 Jahren ($M = 27.11$, $SD = 8.377$), wobei die Studierenden (incl. der Absolventen) im Mittel $M_S = 23.64$ Jahre alt sind (zwischen 20 und 28) und die Dozenten $M_D = 42.4$ (zwischen 33 und 53 Jahren).

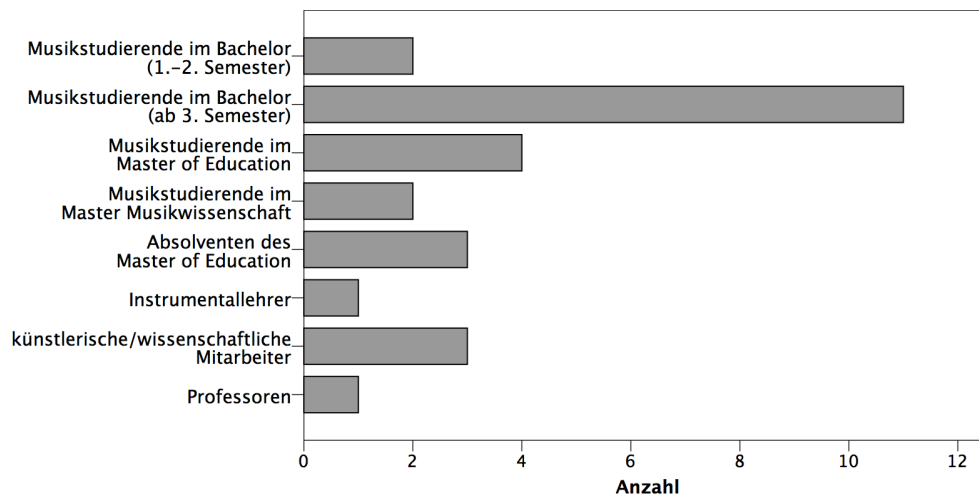


Abbildung 4.1: Zusammensetzung des bewertenden Publikums

Insgesamt sind zwei Drittel der Teilnehmer weiblich, allerdings kehrt sich dieses Ungleichgewicht bei den Dozenten um. Von den fünf Dozenten sind lediglich zwei weiblich, bei den Studierenden und Absolventen hingegen sind es 16 weibliche und nur sechs männliche Teilnehmer. Alle Versuchspersonen sind Universitätsangehörige oder hatten durch ihr Studium hier einen Bezug zur Universität, sodass alle mit den Vorspielmodalitäten einigermaßen vertraut waren, abgesehen ggf. von Studierenden im ersten Semester, die selbst noch nicht vorspielen müssen, und einer Dozentin, die erst im seit Beginn des Wintersemesters hier arbeitet.

4.3 Erhebungsinstrumente

In diesem Abschnitt werden alle Erhebungsinstrumente knapp beschrieben, ohne aber jede einzelne Frage zu nennen. Die Fragebögen sind in vollem Umfang im Anhang dieser Arbeit abgedruckt.

4.3.1 Fragebogen für die Musiker

Im Musikerfragebogen (s.Anhang) wurden zunächst Daten zu Person und Studium erhoben, zur Vorbereitung des Vorspiels, sowie zur Wahrnehmung und zum Umgang mit der Bewertungssituation. Dabei wurde unter anderem gefragt, ob sie sich aktuell mit dem Stück auf eine Prüfung oder einen Wettbewerb vorbereiten, an dem Abend nervöser waren als sonst und ob sie mehr auf ihr Bühnenverhalten geachtet

haben. Die Musiker waren bei der Einführung für das Publikum ebenfalls bereits anwesend und waren insofern informiert, dass nicht bloß klingende Musik sondern die Gesamtperformance bewertet werden sollte. Dabei wurden überwiegend dichotome Fragen, kurze offene Antworten (z. B. für die Semesterzahl) und fünf-stufige Likert-Skalen von „stimme gar nicht zu“ bis „stimme voll zu“ benutzt. Ebenfalls wurden die Musiker gefragt, ob sie mit ihrem Auftritt zufrieden seien, es besser oder schlechter als in den Proben lief und zum Schluss um eine Selbsteinschätzung gebeten. Diese Selbsteinschätzung wurde auf einer kontinuierlichen Skala erhoben, um eine direkte Vergleichbarkeit zu den Echtzeit- und Gesamtwertungsdaten des Publikums zu haben, die in der Papierform als Balken dargestellt wurde, auf dem ein Strich an einer Stelle für die Wertung eingezeichnet werden sollte. Der Balken war genau 10 cm breit um bei der Auswertung später gut ablesen zu können. Es wurde eine Beispiel gezeigt, wie eine Wertung vorgenommen werden soll und danach in zwei Varianten die Bewertung vom Musiker erhoben, einerseits als reine Selbsteinschätzung „Ich würde die Performance unseres Ensembles (bzw. meine Solo-Performance) so bewerten:“ und andererseits als Einschätzung, wie das Publikum wohl gewertet habe: „Ich denke, das Publikum hat die Performance unseres Ensembles (bzw. meine Solo-Performance) so bewertet:“. Zum Schluss wurde noch Platz für Anmerkungen zum Experiment gelassen.

4.3.2 Interface zur Echtzeit-Datenerhebung

Zur Erhebung der Echtzeitdaten kamen 28 iPads mit *emoTouch 3.0* als Interface zum Einsatz. Die iPads waren überwiegend iPad minis der ersten oder zweiten Generation (23 von 28), die übrigen waren „normalgroße“ iPads, mind. dritter Generation. Vier der iPads wurden von den Probanden selbst für den Versuch mitgebracht, mussten allerdings vorab umfassend eingerichtet werden. Die übrigen, universitätseigenen Geräte wurden mit Hilfe des *Apple Konfigurator*s für den Versuch mit *emoTouch* ausgestattet und eingestellt. Dadurch konnte verhindert werden, dass Probanden eigenständig die App verlassen.

emoTouch ist eine iPad-App, die für kontinuierliche Erhebungen in einer oder zwei Dimensionen entwickelt wurde, kann sowohl gleichzeitig als Wiedergabegerät benutzt werden als auch bei Live-Events eingesetzt werden (vgl. Louven & Scholle, in Vorbereitung). Im Experiment wurde die Version 3.0, wie sie derzeit im App Store verfügbar war, verwendet und wurde so konfiguriert, dass den Zuhörern ein horizontal beweglicher Zeiger auf einer kontinuierlichen Skala von *sehr schlecht* bis *sehr gut* zur Bewertung der Performances angeboten wurde (vgl. Abb. 4.2). Der



Abbildung 4.2: emoTouch-Session

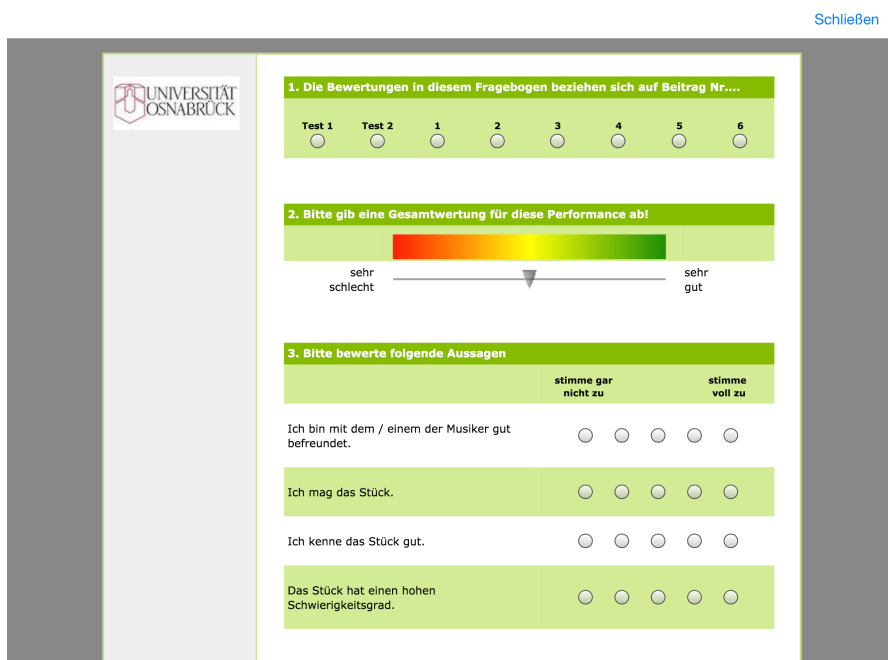


Abbildung 4.3: Webview mit Online-Fragebogen

Zeiger konnte von einer beliebigen Höhe auf dem Display aus gesteuert werden, blieb aber immer auf der Linie und hielt die Position, falls eine Versuchsperson trotz der Anweisung, dies möglichst nicht zu tun, den Finger vorübergehend wegnahm, sodass er von dort aus anschließend weiter bewegt werden konnte. Das iPad spielte in der Live-Situation selbst keine Musik ab und die Versuchspersonen mussten jeden Durchlauf selbst beenden, wenn der Musiker die Bühne verlassen hatte, indem sie den *exit*-Knopf oben links betätigten. Danach öffnete sich ein Fenster mit einem Webbrowser, der automatisch den Fragebogen anzeigte, der aber als einzige eigene Funktion einen Knopf zum beenden hatte (vgl. Abb. 4.3). Danach gelangten sie zurück auf die Startseite, von der aus sie selbstständig, sobald die Moderation begann oder spätestens kurz bevor die Musiker die Bühne betraten, die Sessions laden mussten, um im Erhebungsbildschirm schnell genug auf Start drücken zu können.

Die Skala von *sehr schlecht* bis *sehr gut* wurde ausgewählt und eine Notenskala ausgeschlossen um einerseits eine kontinuierliche Skala zu haben, die nicht suggeriert, dass man den Zeiger auf bestimmte Punkte, z. B. „2.3“ schieben soll, und andererseits auch Deckeneffekten vorbeugt (vgl. Abschnitt 2.4.3). Erfahrungsgemäß ist die Notengebung speziell im instrumentalen Bereich im Institut durchgehend sehr gut, sodass eine 2.3 im Hauptinstrument meist bereits als schlecht empfunden wird, insofern also durch die allgemeinere Skala verhindert werden soll, dass ausschließlich sehr gut bewertet wird und Unterschiede dadurch möglicherweise nicht mehr nachweisbar sind.

4.3.3 Fragebogen für das Publikum

Der Fragebogen für die Zuhörer (s. Anhang) wurde auf der Plattform www.soscisurvey.de erstellt, online verfügbar gemacht und in der Webview aufgerufen. Über die URL konnte der Name der Versuchsperson, der automatisch die Startuhrzeit des Durchgangs enthält, übergeben werden, sodass die Daten aus *emoTouch* und aus dem Fragebogen problemlos der Person und dem gehörten Stück zugeordnet werden konnten, ohne erneut einen Code zur Person abfragen zu müssen. Der Fragebogen ist so strukturiert, dass er sowohl für die kurzen Bewertungspausen zwischen den Stücken funktioniert, als auch am Schluss einmalig nach der Bewertung des Stückes noch personenbezogene Daten erhoben werden können, indem die Versuchspersonen nach jeder Bewertung gefragt werden, ob dies das letzte Stück gewesen sei und je nach Antwort die Aufforderung erhalten, das Fenster zu schließen bzw. zu den übrigen Fragebogenseiten weitergeleitet zu werden. Auf

der ersten Fragebogenseite wurde als zunächst abgefragt, auf welchen Auftritt sich die Bewertung bezieht und dann sollte auf einer kontinuierlichen Skala, ebenfalls von *sehr schlecht* bis *sehr gut* entsprechend der bei der kontinuierlichen Erhebung verwendeten Skala, eine Gesamtbewertung für die Performance abgegeben werden. Die Skala sollte vor allem aufgrund besserer Vergleichbarkeit identisch sein, aber grundsätzlich greifen hier die gleichen Argumente, wie bei dem Interface für die Echtzeit-Erhebung. Graphisch wurde versucht, die Skalen möglichst ähnlich zu gestalten, sodass beide ein graues, nach unten zeigendes Dreieck als Zeiger hatten und die Farbskala identisch war, allerdings war die Breite der Skala im Fragebogen deutlich geringer als in *emoTouch*. Außerdem wurden die Versuchspersonen gefragt, in wie weit sie mit dem bzw. einem der Musiker befreundet seien, ihnen das Stück gefalle, sie das Stück kennen und sie es für schwierig zu spielen halten, um später überprüfen zu können, ob diese Dinge Einfluss auf die Bewertung haben. Speziell wichtig war es, zu erheben, ob die Versuchspersonen Musiker, mit denen sie gut befreundet sind, anders bewerten, da in dem relativ kleinen Institut davon auszugehen war, dass sehr viele sich untereinander kennen. Auf der nächsten Seite kam bereits die Filterfrage, ob dies das letzte Ensemble gewesen sei. Wurde hier „ja“ ausgewählt, gelangten die Probanden zu einer „Pausen“-Seite mit der Anweisung, das iPad bis zum Ende des Vorspiels beiseite zu legen und erst wenn alle übrigen Musiker gespielt haben, auf „weiter“ zu klicken. Dann wurden auf drei weiteren Seiten demographische Daten erhoben, auch zum Studienfortschritt bzw. Position an der Universität und außerdem zur Erfahrung mit solchen Vorspielsituationen, der Bewertungssituation, sowie Instrumentalunterricht als Lehrender. Danach sollten noch einige Items zur Handhabung der App bzw. generell dem Umgang mit der Bewertungsaufgabe auf fünf-stufigen Likert-Skalen bewertet werden. Freiwillig konnten die Probanden am Schluss in einem Feld Anmerkungen zum Experiment hinterlassen, bevor sie auf die Schlussseite gelangten.

4.4 Ablauf

Bereits vor Beginn des eigentlichen Vorspiels wurde eine Einführung für das Publikum durchgeführt. Es wurde die Funktionsweise des iPad-Interfaces erklärt, indem ein Durchlauf ohne Musiker durchgespielt wurde und die Personen sich in Ruhe die Funktionsweise und auch Darstellung der App ansehen konnten inkl. dem Ausfüllen eines Online-Fragebogens. Dabei wurde ihnen zu Beginn auch erklärt, dass die Performance von Auf- bis Abgang bewertet werden sollte und sie wurden

darauf hingewiesen, dass sowohl klingende Musik, als auch das Bühnenverhalten des Musikers zu bewerten sei – entsprechend der Formulierung wie sie auch auf dem iPad zur Erinnerung die gesamte Zeit angezeigt wurde. Die Probanden bekamen etwas Zeit, das Interface in Ruhe auszuprobieren, und nebenbei den Auftrag, den Finger bei der Bewertung zwischendurch nicht wegzunehmen, um nicht zu verrutschen, und auszuprobieren, wie gut das funktioniert, die Bewertung anzupassen, ohne durchgehend auf das iPad zu schauen. Auch beim Fragebogen wurde direkt der Hinweis gegeben, dass bei diesem Durchlauf bei der ersten Frage „Test 1“ auszuwählen sei. Anschließend gab es einen zweiten Testlauf mit Musik, wozu ein Querflötist mit dem kurzen Solostück „Trost“ von Ernesto Köhler aufgetreten ist, und genau so bewertet werden sollte wie später die Musiker in dem Vorspiel, deren Daten ausgewertet wurden. Wieder wurden vor der Anmoderation des Stücks darauf hingewiesen, dass dies Testdurchlauf Nr. 2 sei und dies insofern im Fragebogen auszuwählen sei. Nachdem auch dieser Durchgang abgeschlossen war, wurden noch eine Reihe Anweisungen für den weiteren Verlauf gegeben: Zunächst wurde ein letztes Mal gesagt, wie der erste Frage im Bogen zu behandeln sei („Jetzt Nr. 1 auswählen!“). Mehrere Musiker, die gemeinsam auftreten, sollten als ein Ensemble bewertet werden, was bei den Probedurchgängen noch nicht zum tragen kam. Den Musikern wurde gesagt, sie sollten sich möglichst auf die Performance konzentrieren und soweit nötig auf das eigene iPad, um dezent darauf hinzuweisen, dass man nicht die Bewertungen der Nachbarn beobachten sollte und sich folglich möglichst nicht davon beeinflussen lassen sollte. Zuletzt kam der Hinweis, dass ein Frühstart kein Problem sei, wer also vor Beginn des Auftritts bereits den zweiten Startknopf gedrückt hatte, einfach abwarten konnte und dann gemeinsam mit den anderen mit der eigentlichen Bewertung beginnen.

Zu Beginn des Vorspiels wurde für das übrige Publikum, welches z. T. aber den Probelauf bereits mit gesehen hatte, kurz erläutert, dass in der ersten Hälfte des Vorspiels eine Studie zu kontinuierlichen Performancebewertungen, also Bewertungen während die Musiker spielen, stattfindet und dafür zwischen den Beiträgen eine kurze Bewertungspause gemacht werden muss.

4.5 Beobachtungen im Experiment

Das Experiment verlief im Wesentlichen problemlos und es musste nur in einem Fall nach dem ersten Stück kurz eingegriffen werden, als sich jemand meldete, der den Fragebogen vorzeitig geschlossen hatte. Es funktionierte nicht, dass auf dem Knopf

auf der Startseite „Laden“ angezeigt wurde, was allerdings nicht auf eine falsche Voreinstellung sondern einen Bug in der App zurückzuführen war. Stattdessen stand auf diesem Knopf ebenfalls „Start“, was zunächst für etwas Verwirrung sorgte, weil es anders angesagt wurde, aber zu keinen weiteren Problemen führte. Einige Personen hatten Probleme, den exit-Knopf zu treffen, brauchten dafür mehrere Versuche und einige suchten sogar Hilfe beim Nachbarn. Die vergeblichen Versuche, diesen Knopf zu drücken, führen zwangsläufig dazu, dass dadurch der Zeiger für die Bewertung nach ganz links, wo sich der Knopf auf dem Bildschirm befindet, also an das negative Ende der Skala bewegt wird, was bei der Auswertung berücksichtigt werden und keinesfalls als letzte Bewertung interpretiert werden darf. Bei einigen Personen funktionierte es nicht, dass der Fragebogen am Ende des Vorspiels noch auf der richtigen Seite war und weiter bearbeitet werden konnte, was sowohl an falscher Bedienung wie z. B. vorzeitigem Schließen des Fensters gelegen haben kann, aber auch an Netzwerkproblemen, weil das iPad nicht durchgängig im W-LAN online blieb und die Seite nicht entsprechend neu geladen werden konnte. Für diesen Fall war aber im Fragebogen bereits vorgesorgt, indem die Nummerierung des zu bewertenden Auftritts bis sechs ging und beim Auftreten des Problems die Anweisung gegeben wurde, einen weiteren Durchgang mit ganz kurzer Echtzeitbewertung für einen imaginären sechsten Musiker zu starten, dem eine beliebige Gesamtbewertung zu geben, sodass man wieder auf die entsprechende Fragebogenseite gelangte, um die übrigen Fragen zu beantworten.

5 Ergebnisse

Die Datenauswertung erfolgt gegliedert nach den drei zentralen Forschungsfragen. Entsprechend beschäftigt sich der erste Abschnitt ausschließlich mit den Ergebnissen aus den Fragebögen, vorrangig mit den Gesamtbewertungen des Publikums und möglichen Einflussfaktoren auf diese Bewertungen. Darüber hinaus werden diese mit den Einschätzungen der Musiker selbst verglichen. Im zweiten Abschnitt werden explorativ die Zeitreihendaten betrachtet und Auswertungsmöglichkeiten und Ergebnisse vorgestellt, wobei die Analyse teilweise nur exemplarisch durchgeführt wird. Im dritten Abschnitt werden dann die kontinuierlichen Bewertungen mit den retrospektiven in Bezug gesetzt und es wird untersucht, ob und wie gut die in Abschnitt 2.6 vorgestellten Modelle den Zusammenhang erklären können.

5.1 Konsistenz der Gesamtbewertungen

5.1.1 Durchschnittliche Gesamtbewertungen und Streuung

Bei den Gesamtbewertungen wurde von den Versuchspersonen die gesamte angebotene Skala ausgenutzt, die intern von -100 bis +100 kodiert wurde, wobei tendenziell eher in der oberen Skalenhälfte gewertet wurde. Beim Vergleich der mittleren Bewertungen (s. Abb. 5.1) ist erkennbar, dass Ensemble fünf knapp am besten bewertet wurde ($M = 80.96$), dicht gefolgt von Ensemble bzw. in diesem Fall besser Musikerin vier ($M = 72.48$). Jeweils ist die Streuung der Bewertungen mit $SD \leq 18$ deutlich geringer als bei den übrigen Performances. In der Mitte der Rangordnung landet Ensemble eins, auch bezüglich der Standardabweichung belegt es den dritten Platz ($M = 42.08$, $SD = 24.839$). Die Plätze vier und fünf mit deutlichem Abstand zu den ersten dreien und einer wesentlich größeren Streuung belegen Ensemble zwei ($M = -6.60$, $SD = 42.429$) und drei ($M = -31.08$, $SD = 35.232$). Auch wenn das Ensemble mit der zweitschlechtesten Gesamtbewertung mit der größten Streuung bewertet wurde und die niedrigste Standardabweichung beim Ensemble bzw. der Musikerin mit der zweitbesten Bewertung vorliegt, ist das Publikum sich augenscheinlich bei den schlechteren Performances deutlich uneiniger, während gute Performances relativ einstimmig als solche wahrgenommen werden. Dabei ist

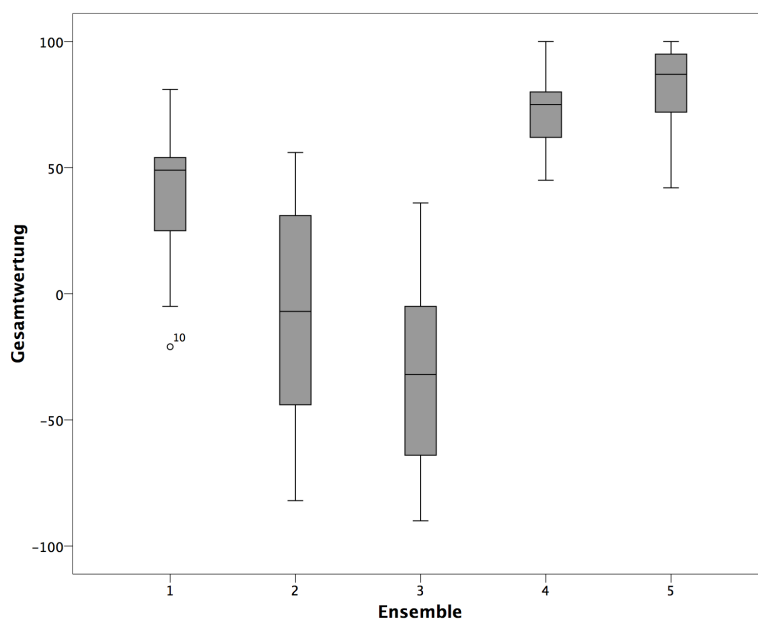


Abbildung 5.1: Bewertungen der fünf Performances im Vergleich

die Einordnung als „gute“ oder „schlechte“ Performance lediglich an den Publikumswertungen abgelesen. Dieser Eindruck lässt sich durch die Berechnung einer Pearson-Korrelation zwischen Mittelwert und Standardabweichung bestätigen und mit $r(3) = -.903$, $p = .036$ liegt hier sogar ein starker negativer Zusammenhang vor.

Die Interpretation, nur für gute Performances sei sich das Publikum einig, liegt nahe, allerdings sollte berücksichtigt werden, dass bei den beiden bestbewerteten Ensembles eine mittlere Bewertung so kurz vor dem oberen Ende der Skala ist, dass entsprechend große Abweichungen einzelner Personen gegenüber dem Mittel nach oben gar nicht möglich sind. Bei Ensemble fünf beträgt der Median 87.00 und ist damit 6.04 höher als das arithmetische Mittel, sodass bis zum Skalenende weniger als eine Standardabweichung Abstand ist. Der Modalwert von 100 entspricht der höchstmögliche Wertung (vgl. Abb. 5.2). Daraus resultiert eine Schiefe von $-.915$, was ebenfalls die Annahme eines Deckeneffekts unterstützt. Allerdings trifft diese Problematik bei der vierten Performance nur abgeschwächt zu und die noch geringere Streuung lässt sich nicht einfach auf die Skala zurückführen. Der Median liegt mit 2.52 Distanz nur geringfügig über dem Mittelwert, der Modus nochmals etwas höher und die Schiefe ist mit $-.185$ vernachlässigbar. Bei Ensemble zwei

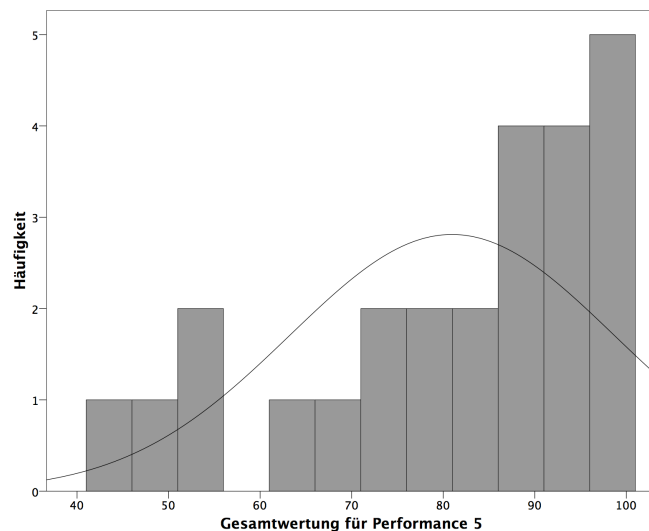


Abbildung 5.2: Verteilung der Wertungen für Ensemble 5

liegt offenbar keine Normalverteilung vor, sondern es gibt zwei Gruppen, die unterschiedlich gut bewerten (vgl. Abb. 5.3).

5.1.2 Konsistenz der Bewertungen

Für die Versuchsteilnehmer gab es keinerlei Erklärung zu der Skala und Erfahrungswerte, wie sie bei einer Notenskala vorhanden gewesen wären, existieren ebenfalls nicht. Somit ist eine sehr unterschiedliche Nutzung der Skala möglich, bei der jeder sowohl selbst entscheidet, wie hoch die Messlatte für eine sehr gute Bewertung hängt, als auch welchen Niveau-Unterschied ein bestimmter Abstand auf der Skala ausmacht. Dementsprechend ist eine relative Auswertung, bei der weniger die absoluten Wertungen berücksichtigt werden, dafür aber, welchen Spielraum der Skala die Probanden ausgenutzt haben, sinnvoller.

Von den Auswertungsmethoden aus der Literatur kann damit das Zählen von Meinungsverschiedenheiten, wie bei Smith (2004) angewandt, bereits ausgeschlossen werden. Auch die viel größere Anzahl von Juroren ist problematisch, wenn das Maß für Übereinstimmung sich lediglich an der maximalen Differenz zwischen zwei Ratern pro Stück orientiert. Überträgt man die Definition von „disagreement“, von einem bzw. zwei ganzen Punkten auf einer fünfstufigen Skala auf die in dieser Studie verwendete kontinuierliche Skala von -100 bis 100, müsste man es als Abweichung zählen, wenn zwei beliebige Zuhörer um 50 bzw. 100 verschieden werten. Die Ranges

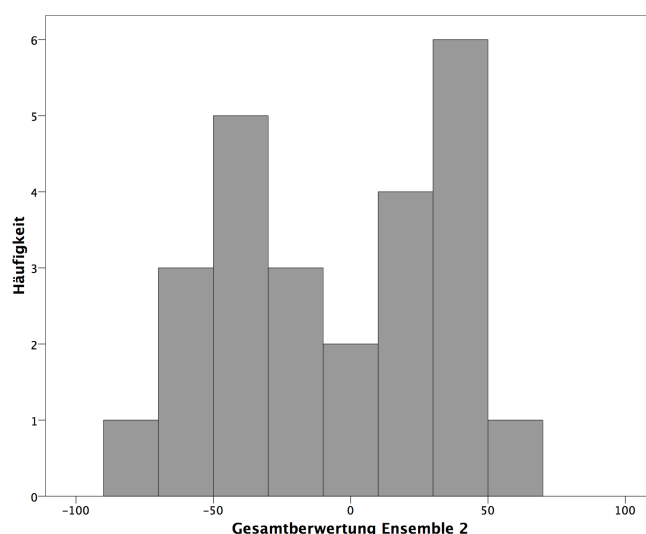


Abbildung 5.3: Verteilung der Wertungen für Ensemble 2

der absoluten Punkte bei den Stücken sind in Auftrittsreihenfolge 102, 138, 126, 55 und 58, was nach Smiths Definition 60 % bzw. 100 % Meinungsverschiedenheiten bedeuten würde.

Bedeutend sinnvoller ist das Berechnen von Inter-Rater-Korrelationen, weil diese unabhängiger von den absoluten Skalenwerten sind und ebenfalls eine Vergleichbarkeit mit den Ergebnissen aus der Literatur gewährleisten. Da eine Korrelationsmatrix für 27 Rater sehr unübersichtlich wird, werden nur die Rahmendaten maximale sowie minimale und die mittlere Korrelation, welche alle einzelnen Werte berücksichtigt, betrachtet. Dabei ergeben sich z. T. sehr hohe relative Übereinstimmungen mit Korrelationskoeffizienten von $r(3) > .999$ und selbst die beiden am schlechtesten übereinstimmenden Rater kommen noch auf eine Korrelation von $r(3) = .589$, was im Vergleich zu den Ergebnissen von Davidson und Coimbra (2001) und Smith (2004) ein deutlich stärkerer Zusammenhang ist (s. Tabelle 5.1). Aussagekräftiger ist allerdings speziell bei der größeren Zahl der Bewertenden in dieser Studie die mittlere Inter-Rater-Korrelation von $.883$, die sich auch bei Davidson und Coimbra (2001, S. 39) aus der angegebenen Korrelationstabelle berechnen ließ und mit $.602$ deutlich schwächer ist. Werden aus der vorliegenden Studie nur die Dozentenbewertungen ausgewertet, um eine mit diesen Ergebnissen vergleichbare Anzahl Juroren mit vergleichbarer Position an der Universität zu haben, ändert sich bei der Gegenüberstellung wenig. Die Range der einzelnen Korrelationen wird geringer, sodass die niedrigste noch deutlich höher liegt und die mittlere Korrelation

	Dozenten	Alle	Davidson und Coimbra (2001)	Smith (2004)
Ensembles	5	5	21	78
Rater	5	24	4	5
mittlere Inter-Rater-Korrelation	.847	.883	.602	unbekannt
min. Korrelation	.715	.589	.4 (nicht signifikant)	.48
max. Korrelation	.999	>.999	.71	.78
Cronbach's α	.945	.990	unbekannt	.9

Tabelle 5.1: Vergleich der Inter-Rater-Reliabilität

geringfügig niedriger ist. Für einen Vergleich zu Smith (2004) funktioniert dieses Kriterium nicht, aber anstelle dessen wurde in seiner Studie ein Cronbach's α von .9 für die Inter-Rater-Reliabilität angegeben. Trotz des bereits exzellenten Wertes, misst das Cronbach's α in dieser Studie sogar eine noch höhere Reliabilität von .990. Dabei bewerteten die gesamte Gruppe und die Dozenten vergleichbar konsistent, wobei die Gruppe im Ganzen eine etwas höhere mittlere Inter-Rater-Korrelation erreicht, während die niedrigste Korrelation bei den Dozenten bedeutend höher ist. Insofern unterstützen die Ergebnisse die beiden diesbezüglichen Hypothesen 1a und 1b.

5.1.3 Einfluss von Freundschaftsgrad, Präferenz, Bekanntheit und eingeschätztem Schwierigkeitsgrad

Obwohl die Parameter Freundschaftsgrad mit dem oder einem der Musiker, Präferenz, Bekanntheit des Stücks, sowie eingeschätzter Schwierigkeitsgrad erhoben wurden, um zu überprüfen, inwiefern diese Einfluss auf die Bewertung haben, beispielsweise indem gute Freunde systematisch besser bewertet werden, lässt sich hier keine Kausalität ableiten. Es wurden lediglich Zusammenhänge in Form von Korrelationen unter Einbezug aller Bewertungen für alle Performances untersucht, wobei die Präferenz für das Stück die höchste Korrelation mit der Gesamtwertung aufweist ($r(125) = .717, p < .001$). Ein mittlerer Zusammenhang ergibt sich zwischen eingeschätztem Schwierigkeitsgrad und der Gesamtwertung ($r(125) = .521, p < .001$), während der Freundschaftsgrad zwar signifikant, aber nur schwach mit der Bewertung korreliert ($r(125) = .208, p = .019$) und, ob der Zuhörer das Stück kennt, überhaupt nicht ($r(125) = .053$). Allerdings gibt es zwei mögliche Interpre-

Ensemble	1	2	3	4	5
Pearson-Korrelation	.496	-.031	.781	.212	-.114
Signifikanz	.010	.883	≤ .001	.308	.587
N	26	25	26	25	25

Tabelle 5.2: Korrelationen zwischen der Gesamtbewertung und der Präferenz für das Stück

tationen für diese Korrelationen, die im Folgenden am Beispiel *Präferenz* erläutert werden. An der Gesamtkorrelation ist nicht erkennbar, inwiefern lediglich insgesamt Stücke, die den Zuhörern schlechter gefielen, auch schlechter bewertet wurden, weil sie die Stücke nicht mochten, oder ob zufällig diese Stücke schlechter gespielt worden sind. Möglicherweise wurden also die Stücke zwei und drei deutlich schlechter bewertet, weil das Publikum Barockmusik im Vergleich zu den übrigen Darbietungen langweilig fand. Alternativ könnte das Publikum das Stück als schlecht gespielt wahrgenommen haben und es gefiel der Mehrheit unabhängig davon nicht. An dieser Stelle ist es schwierig, das eine oder andere Phänomen anhand der Daten zu belegen. Zudem ist nicht auszuschließen, dass das Präferenzurteil sehr stark von der Performance-Qualität abhängt und nicht umgekehrt, weil die Stücke überwiegend unbekannt waren und ein einmalig schlecht dargebotenes Stück dem Publikum wahrscheinlich weniger gefallen wird. Insofern wichtiger als die Gesamtkorrelation ist die Frage, ob die Personen, denen ein bestimmtes Stück gefallen hat, dieses besser bewertet haben, als diejenigen, denen es weniger gefallen hat. Für eine allgemeine Aussage darüber, die alle Stücke miteinbezieht, muss daher eine andere Zusammenrechnung der Korrelationen mit den möglichen Einflussfaktoren erfolgen, die nicht von der Bewertungstendenz für eine Performance abhängig ist.

Bei der Einzelbetrachtung der Stücke (s. Tabelle 5.2) fällt auf, dass nur bei Performance drei eine vergleichbar hohe Korrelation erreicht wird wie im Gesamtbild und darüber hinaus nur noch bei Ensemble eins sich überhaupt ein signifikanter Effekt zeigt. Für die schwache Korrelation bei Performance 4 beträgt die Teststärke nichtmal 20%, sodass keine sichere Aussage getroffen werden kann. Der starke Zusammenhang, den die übergreifende Korrelation über alle Stücke suggeriert, lässt sich deshalb nicht auf die einzelnen Stücke übertragen und muss daher zu großen Teilen auf Unterschiede zwischen den Performances bzw. der Präferenz für die unterschiedlichen Stücke zurückgeführt werden. Ein Stück weit lässt sich dieses Ergebnis durch die geringe Varianz der Präferenzurteile begründen. Speziell für Ensemble zwei ist dies nicht aufschlussreich, da trotz der großen Streuung sowohl

der Performancebewertungen als auch der Gefallensurteile hier überhaupt kein Effekt zu finden ist. Was den eingeschätzten Schwierigkeitsgrad betrifft, ergibt sich ein einheitlicheres Bild mit Korrelationen von $r = .199$ bis $.556$. Die größten drei davon sind signifikant, doch auch hier ist der Effekt kleiner als im Gesamtbild. Immerhin gibt es hier einen Zusammenhang mit einer Korrelation von $r(23) = .503$, $p = .010$ bei Ensemble zwei und auch die Bekanntheit des Stückes steht bei dieser Performance im Zusammenhang mit der Bewertung, und zwar negativ mit $r(23) = -.513$, $p = .009$, sodass hiermit ein Teil der großen Streuung erklärt werden kann. Im Scatterplot (Abb. 5.4) lässt sich erkennen, dass einige wenige, die das Stück kennen, bedeutend schlechter werten, als die Mehrheit, denen das Stück gänzlich unbekannt ist. Mit einem der Musiker befreundet zu sein, geht für Performance eins und drei mit einer etwas besseren Bewertung einher ($r(24) = .393$, $p = .047$ bzw. $r(24) = .341$, $p = .088$), während in den übrigen Fällen max. knapp ein kleiner Effekt erreicht wird, der auch bei weitem nicht signifikant wird.

Um ein Gesamtbild ohne Einflüsse der unterschiedlichen Bewertung der einzelnen Stücke zu erhalten, bei dem aber alle Daten mit ausgewertet werden, wurden die abschließenden Urteile jeweils um minus den Mittelwert verschoben. Diese Verschiebung ist vergleichbar mit einer z-Transformation, die nur zur Hälfte durchgeführt wird. Somit ist die mittlere Bewertung für jedes Stück nach der Verschiebung 0, allerdings wurde die Streuung nicht normiert und die Standardabweichung bleibt erhalten. Diese Variante wurde gewählt, um die Mittelwertunterschiede auszugleichen ohne die Abweichungen der einzelnen Personen davon zu beeinflussen, damit nicht eine geringe Streuung bei einem Stück zu sehr verstärkt wird, obwohl kaum Unterschiede vorlagen. Auf der anderen Seite bleiben auch große Abweichungen von Probanden von der mittleren Bewertung des Stückes erhalten. Mit dieser Skala wurden erneut Korrelationen mit der Gesamtbewertung berechnet, bei denen sich in drei Fällen ein signifikanter Zusammenhang ergibt. Dabei sind allerdings die Korrelationskoeffizienten klein. Der größte ergibt sich für den Schwierigkeitsgrad mit $r(125) = .291$ ($p = .001$) und alle anderen sind kleiner als $.2$, wobei der Zusammenhang mit dem Freundschaftsgrad nicht signifikant wird.

Das Publikum bezieht offenbar den Schwierigkeitsgrad in die Bewertung durchaus ein, wenn auch in unterschiedlichem Umfang je nach Stück. Die Bekanntheit des Stückes hingegen ist insgesamt unbedeutend und der Freundschaftsgrad spielt im Gesamtbild auch eine untergeordnete Rolle. Zusammenfassend lässt sich festhalten, dass die Zusammenhänge bei den Stücken allerdings sehr unterschiedlich sind und sich daher nur schwer allgemeiner Aussagen treffen lassen.

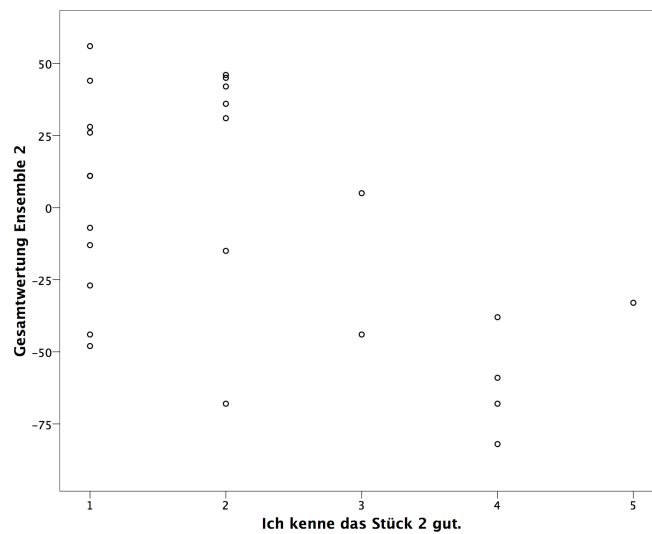


Abbildung 5.4: Streudiagramm: Bewertung und Bekanntheitsgrad bei Stück 2

5.1.4 Unterschiede zwischen Studierenden und Dozenten sowie zwischen den Geschlechtern

Dieser Abschnitt beschäftigt sich damit, ob Männer und Frauen bzw. Dozenten und Studierende grundsätzlich besser oder schlechter bewerten. Ein t-Test für alle Gesamtwertungen aller Stücke für beide Geschlechter im Vergleich wird trotz Mittelwertunterschied von 35.58 zu 21.88 nicht signifikant ($t(125) = 1.385$, $p = .169$). Unter Berücksichtigung der Standardabweichungen entspräche dies einem kleinen Effekt ($d = .271$) und die Test-Power ist sehr gering ($1 - \beta = .294$). Für jedes einzelne Stück berechnet, liefert ein die Geschlechter vergleichender t-Test hingegen bedeutend interessantere Ergebnisse. Für Ensemble drei wird der Effekt fast signifikant ($t(24) = 1.995$, $p = .057$, $d = .858$, $1 - \beta = .483$) und für Ensemble vier signifikant mit $p = .015$ ($t(23) = 2.623$, $d = 1.221$, $1 - \beta = .778$) und das trotz oder genau wegen der im Vergleich zu den übrigen Stücken geringen Varianz. Bei den übrigen Ensembles wird der t-Test nicht signifikant, aber mit einer größeren Stichprobe hätte man möglicherweise Effekte finden können. Die Effektstärken, sofern ein Effekt existiert, betragen für Ensemble eins $d = .040$, für Ensemble zwei $d = .691$ und für Ensemble fünf $d = .230$ und sind im Mittel größer als der Effekt für alle Ensembles zusammen berechnet.

Auch bei den Unterschieden zwischen Dozenten und Studierenden (Absolventen hier mit inbegriffen) tritt das gleiche Problem auf. Für alle Stücke wird ein t-Test

zwischen diesen beiden Gruppen nicht signifikant ($t(125) = 1.569$, $p = .119$), wobei zu berücksichtigen ist, dass die Vergleichsgruppen mit 102 Studierenden- bzw. Absolventen- und 25 Dozentenwertungen sehr unterschiedlich groß sind. Es ergibt sich insgesamt ein Effekt von $d = .352$ und eine Teststärke von $1 - \beta = .347$ dafür, dass die Dozenten schlechter bewerten, sodass nicht auszuschließend ist, dass der Effekt existiert und lediglich hier nicht belegt werden konnte. Bei einer Einzelbetrachtung der Performances wird der t-Test auch für den Experten-Studenten-Vergleich für Ensemble eins ($t(24) = 2.583$, $p = .016$), drei ($t(24) = 2.067$, $p = .05$) und vier ($t(23) = 2.206$, $p = .038$) signifikant, bei jeweils fünf Dozentenwertungen und 20 bis 21 Studierenden-/Absolventenurteilen. Dabei kommen riesige Effektstärken von $1.035 \leq d \leq 1.305$ zustande, deren Verallgemeinerbarkeit aufgrund der kleinen Gruppen, speziell der fünf Dozenten, allerdings stark eingeschränkt ist. Spannend ist, dass das letzte Stück, welches im Programm gewissermaßen eine Sonderstellung einnimmt, von den Dozenten dem Mittelwert nach überhaupt nicht schlechter bewertet wurde ($M_S = 80.95$, $M_D = 81.00$, $d = .057$), alle anderen aber durchaus, am nächstschwächsten Ensemble 2 mit einer Effektstärke, falls der Effekt existiert, von $d = .404$. Eine noch feinere Abstufung über den Ausbildungsstand, so wie dieser erhoben wurde, mit diversen Kategorien für Studierende wie Dozenten, ist nicht sinnvoll, weil die Fallzahlen pro Gruppe dadurch zu klein werden und selbst bei nur zwei Gruppen bereits problematisch sind.

Aufgrund der Mittelwertunterschiede und insbesondere der sehr verschieden großen Standardabweichungen gehen die Effekte, die sich für einzelne Stücke trotz der geringen Fallzahl zeigen, im Gesamtbild größtenteils verloren. Andererseits sind t-Tests bei Gruppengrößen von 5 zu 20 grundsätzlich problematisch, vor allem dann wenn eine Normalverteilung nicht gegeben ist (vgl. Abb. 5.2 und 5.3 auf Seite 60 und auf Seite 61) oder zusätzlich die Varianzen verschieden groß sind (vgl. Bortz & Schuster, 2010, 122f.). Die Bewertungen von fünf Dozenten und auch die der Studierenden sind ohnehin kaum verallgemeinerbar. Zudem dürften die Effekte für den Geschlechtervergleich bzw. den Dozenten/Studierende Vergleich nicht völlig unabhängig voneinander sein, weil die Geschlechterverhältnisse in den beiden Gruppen umgekehrt sind.

Als Lösungsansatz ist an dieser Stelle wieder auf die verschobenen Bewertungen zurückzugreifen, bei denen jedes Stück im Mittel mit 0 bewertet wurde. Dies hat den Vorteil, dass die Gruppen größer sind, also die Voraussetzungen für den t-Test besser erfüllt sind, aber gleichzeitig systematische Unterschiede dadurch, dass die Stücke unterschiedlich bewertet wurden, herausgefiltert sind. Für den

Unterschied zwischen den Geschlechter ergibt sich so ein signifikanter Unterschied ($t(125) = 2.807, p = .006$) mit mittlerer Effektstärke ($d=.53$). Auch der Unterschied zwischen Dozenten und Studierenden ist so belegbar ($t(125)=3.062, p=.003$) mit einem noch größeren Effekt von $d=.68$. Es bleibt zu bedenken, dass bei dieser Methode die Stücke mit größerer Streuung stärker ins Gewicht fallen, aber wendet man statt der Verschiebung eine z-Transformation an, verhalten sich die Ergebnisse ähnlich. Da die z-Transformation genau das Gegenteil zu Folge hat, nämlich die kleinen Abweichungen zu strecken und ihnen mehr Ausschlagkraft zu geben, kann davon ausgegangen werden, dass der Effekt vorrangig durch das Herausrechnen der unterschiedlichen Mittelwerte nachweisbar wird.

Es folgt, dass die Dozenten etwas schlechter werteten als die Studierenden und Männer schlechter als Frauen. Dies ist bereits ein Teilergebnis bzgl. der Hypothese 1c zur Expertise, da davon auszugehen ist, dass die Hochschuldozenten deutlich mehr Erfahrung im Bewerten von Performances haben. Allerdings wurde die Selbsteinschätzung zur Expertise noch nicht betrachtet und daher an dieser Stelle noch keine Entscheidung getroffen, ob die Hypothese angenommen wird. Es lässt sich darüber hinaus festhalten, dass sich für verschiedene Performances sehr unterschiedliche Effekte zeigen, die, wenn wie bei anderen Studien ausschließlich Klavieraufnahmen ähnlicher Stilistik zum Einsatz kommen, nicht gefunden werden können. Gerade dies zeigt, dass Ergebnisse aus solchen Studien, nur sehr wenig verallgemeinerbar für andere Stücke sind.

5.1.5 Expertise als Einflussgröße

Expertise wurde im Fragebogen einerseits als Selbsteinschätzung erhoben („Meine Erfahrung mit der Bewertung von Performances / Instrumentalunterricht (als Lehrender) schätze ich ein als:“ auf einer Skala von „sehr gering“ bis „sehr hoch“) und andererseits wurde abgefragt, wie oft die Versuchspersonen in den vergangenen zwölf Monaten an vergleichbaren Situationen teilgenommen hatten, nochmals differenziert zwischen der Teilnahme als Musiker und als Zuhörer. Zusammenhänge zwischen diesen Parametern und der Gesamtbewertung finden sich allerdings kaum. Die betragsmäßig größte Korrelation ergibt sich zwischen der Gesamtbewertung und der Selbsteinschätzung der Bewertungserfahrung, ist allerdings mit $r(125) = -.125, p = .162$ nicht signifikant und so gering, dass hier von keinem Zusammenhang auszugehen ist. Alle übrigen Korrelationskoeffizienten haben einen Betrag kleiner 0.1, wobei bei den Ordinalskalen zur Häufigkeit der Teilnahme an vergleichbaren Situationen ein nichtparametrisches Spearman- ρ anstelle der Pearson-Korrelation

berechnet wurde. Betrachtet man an diese Stelle wieder die pro Stück um die mittlere Bewertung verschobenen Variablen, ergeben sich etwas größere Korrelationskoeffizienten und für die Bewertungserfahrung ($r(125) = -.252, p = .004$) sowie die Teilnahme als Zuhörer ($\rho(125) = -.189, p = .034$) jeweils ein kleiner Effekt. Auf die generelle Problematik dieser verschobenen Bewertungsskala wurde bereits hingewiesen, aber in diesem Fall ist interessant, dass sich jeweils eine negative Korrelation ergibt, auch wenn sie relativ gering ist. Je mehr Erfahrung mit der Bewertung jemand hat bzw. ja öfter er oder sie als Zuhörer an Vorspielen teilnimmt, umso schlechter bewertet diese Person. Dieses Ergebnis passt insofern dazu, dass die Dozenten tendenziell schlechter bewerten, da sie Instrumentalprüfungen abnehmen müssen oder oft bei den institutsinternen Studierendenvorspielen anwesend sind, auch wenn nicht auszuschließen ist, dass es Studierende gibt, die häufig als Zuhörer dabei sind oder in ihrer Freizeit als Juror in der Laienmusik tätig sind. Möglicherweise ist der Dozentenstatus ein besserer Indikator für Bewertungsexpertise als die abgefragte Selbsteinschätzung, sodass der Effekt hier im Vergleich klein bleibt. Dennoch sprechen beide Ergebnisse für einen negativen Zusammenhang zwischen Expertise und Bewertungsverhalten und daher wird die Hypothese 1c, dass Teilnehmer mit mehr Bewertungsexpertise kritischer Bewerten, angenommen.

5.1.6 Vergleich zur Selbsteinschätzung der Musiker

Die Selbsteinschätzung der Musiker, sowie ihre erwartete Publikumsbewertung fielen deutlich neutraler aus, als die mittleren Publikumsbewertungen für die Auftritte. Die Selbsteinschätzung, die sich nicht nur auf die eigene Performance bezog, sondern auf die des Ensembles insgesamt, betrug im Mittel 6.29 ($SD = 20.26$) bei einer Range von -24.50 bis 32.50. Die erwartete Publikumsbewertung hatte mit $M = 6.04$ einen sehr ähnlichen Mittelwert, allerdings waren der Median der mit einem Wert von 13.00 auf eine schiefe Verteilung hindeutet, sowie die Standardabweichung von 25.36 hier deutlich höher. Auch die Range von -33.50 bis 34.50 war hier größer. Bei diesen deskriptiven Statistiken bleibt allerdings zu berücksichtigen, dass bis zu vier Wertungen für ein Ensemble vorliegen, und deren Performance folglich mehr ins Gewicht fällt, als die der Solo-Pianistin.

Für den Vergleich mit der Publikumswertung wurden die Differenzen aus den Selbsteinschätzungen bzw. erwarteten Bewertungen und der mittleren Publikumsbewertung für den jeweiligen Auftritt gebildet. Dabei ergibt sich, dass lediglich drei der zwölf Musiker ihren Auftritt tendenziell besser eingeschätzt haben, als das Publikum ihn sah. Die größte Überschätzung mit 32.08 für die Selbsteinschätzung

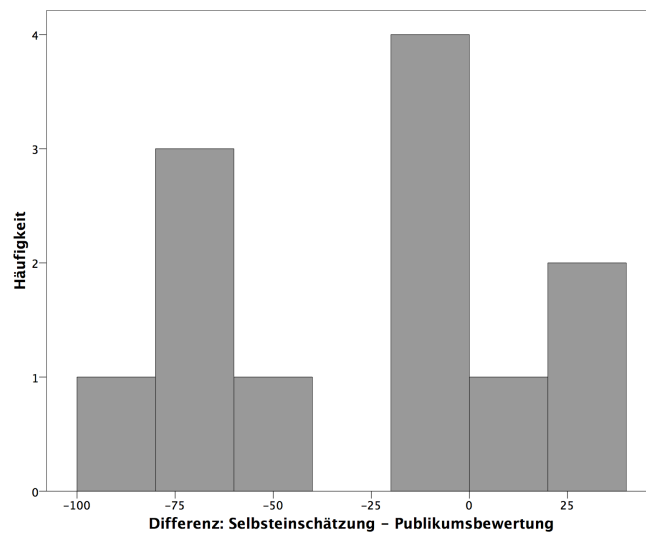


Abbildung 5.5: Histogramm: Bewertungsdifferenzen

bzw. 29.60 für die erwartete Publikumsbewertung ist zwar deutlich, stellt allerdings auf einer 200-stufigen Skala noch keinen tatsächlich großen Bewertungsunterschied dar (vgl. Abb. 5.5). Hingegen unterschätzen deutlich mehr Musiker, wie gut das Publikum sie bewerten würde bzw. bewerteten ihre Performance selbst schlechter, wobei hier in einigen Fällen die Bewertungsdifferenzen bedeutend größer waren. Die größten Bewertungsunterschiede von ca. 60 bis 80 ergaben sich für die letzten beiden Performances, die vom Publikum besonders gut bewertet wurden, während die Musiker sich, wie die übrigen Musiker auch, eher im mittleren Bereich der Skala sahen.

Im Streudiagramm (Abb. 5.6) zeigt sich, dass ein großer Zusammenhang zwischen der Bewertung vom Publikum und den Bewertungsunterschieden zwischen Publikum und Selbsteinschätzung vorliegt, der von einer hohen Korrelation von $r(10) = .901$ ($p \leq .001$) bestätigt wird. Zwischen der Abweichung der erwarteten Publikumsbewertung und der tatsächlichen ergibt sich eine annähernd gleich hohe Korrelation ($r(10) = .848$, $p \leq .001$). Dieser statistische Wert, der trotz sehr kleiner Stichprobe signifikant wird, sollte jedoch nicht überbewertet und auf andere Situationen übertragen werden, gerade weil größere Ensembles hier stärker ins Gewicht fallen und auch der Bewertungsprozess für das Ensemble sicher von der reinen Selbsteinschätzung bei einem Soloauftritt abweicht. Daher kann nur festgehalten werden, dass sich in dieser Stichprobe besser bewertete Ensembles stärker

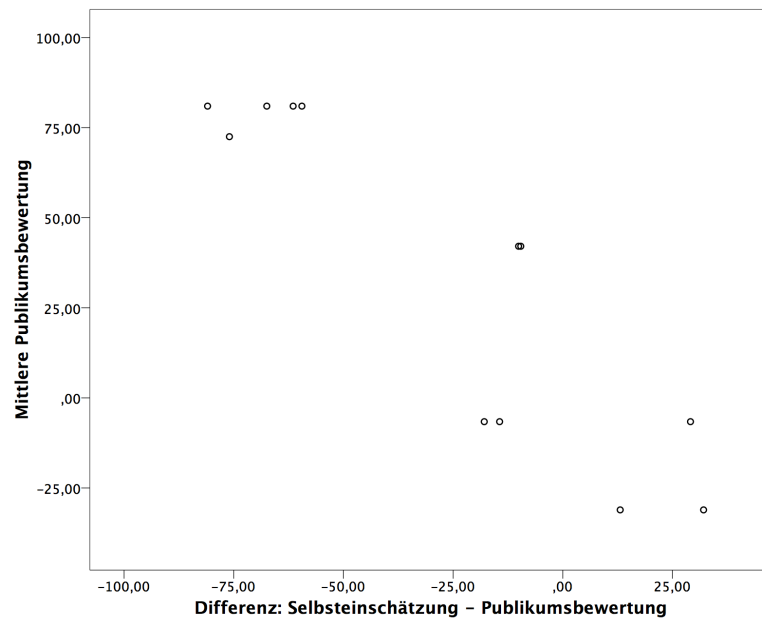


Abbildung 5.6: Streudiagramm: Publikumsbewertung und Abweichung der Selbsteinschätzung von der Publikumsbewertung

unterschätzten im Vergleich zur Publikumsbewertung als schlechter bewertete, die teilweise sich selbst auch etwas besser bewerteten, als die Zuhörer.

Trotz dieser z. T. großen absoluten Unterschiede bei der Bewertung ergeben sich mittlere bis hohe Korrelationen zwischen der mittleren Publikumsbewertung und der Selbsteinschätzung der Musiker ($r(10) = .486$, $p = .109$), bzw. der erwarteten Publikumsbewertung ($r(10) = .714$, $p = .009$). Folglich sind die Bewertungstendenzen durchaus ähnlich, nur die Benutzung der Skala bzw. der gestellte Anspruch unterscheiden. Allerdings hängt die Selbsteinschätzung deutlich schwächer mit der Gesamtbewertung zusammen als die erwartete Publikumsbewertung, was zeigt, dass die Musiker ein Gefühl dafür haben, in welche Richtung das Publikum abweichen wird. Möglicherweise haben manche Fehler sie selbst gestört, sie vermuten aber, dass das Publikum diese nicht oder als weniger störend wahrnimmt. Vermutlich wird die Evaluation der eigenen Leistung auch stärker von den persönlichen Erwartungen, beispielsweise der Erfahrung, wie gut das Stück in den Proben lief, beeinflusst und weicht daher stärker von der Publikumsbewertung ab, das diesen Vergleichspunkt als Einschätzung des Leistungspotenzials nicht hat. Auch wenn die Korrelation mit der Selbsteinschätzung für die geringe Stichprobengröße nicht signifikant wird, lässt sich die Hypothese 1d, die Selbsteinschätzungen seien nicht konsistent mit

den Publikumsbewertungen, verwerfen. Die Selbsteinschätzungen sind somit nicht unabhängig von der von außen wahrgenommenen Leistung. Stattdessen sorgen eher Bescheidenheit oder Unterschätzung für insgesamt schlechtere Bewertungen. Möglicherweise ist das Bewertungsverhalten der Musiker auch als Vorsicht zu verstehen, da die Bewertungen alle sehr in der Mitte der Skala sind, vielleicht um sich nicht zu sehr zu verschätzen. Die Überlegung, dass sich selbst zu überschätzen gewissermaßen peinlich wäre, mag auch zu den niedrigeren bzw. neutralen Bewertungen beigetragen haben, da bei den Musikern Anonymität nur sehr eingeschränkt gewährleistet ist und die Fragebögen eindeutig den Ensembles und in vielen Fällen – aufgrund von Geschlecht oder Semesterzahl – auch den Musikern zuzuordnen sind. Speziell die Pianistin, die als viertes gespielt hat, könnte diesen Gedanken gehabt haben, da sie ihre Selbsteinschätzung sowie die erwartete Publikumsbewertung, nachdem sie einen ersten Eintrag gemacht hatte, noch nach unten korrigiert hat. An dieser Stelle ist die Tendenz der Selbsteinschätzungen für diese Stichprobe recht anschaulich, allerdings müsste für eine aussagekräftigere Antwort eine größere Anzahl Musiker befragt und von einem Publikum oder einer Jury bewertet werden. Außerdem müsste mehr Anonymität gegeben sein, beispielsweise dadurch, dass sich die Beteiligten untereinander schlechter kennen, sodass möglicherweise ehrlichere Einschätzungen abgegeben werden.

5.2 Analyse der Bewertungsverläufe

In diesem Abschnitt werden die kontinuierlichen Performancebewertungen zunächst losgelöst von den Abschlussbewertungen und allen weiteren Angaben aus den Fragebögen betrachtet. Sie werden auf Konsistenz untersucht, wobei absolute Bewertungsunterschiede zu den einzelnen Zeitpunkten sowie Ähnlichkeit der Bewertungsverläufe berücksichtigt werden. Bei den Verläufen liegt der Fokus auf den Bewertungsänderungen und nicht darauf, inwiefern die Bewertungen vergleichbar gut sind. Außerdem werden die Zeitreihen in Verbindung zu den Aufnahmen der Auftritte gebracht, sodass geprüft werden kann, ob ein direkter Zusammenhang zwischen Performance und Bewertung besteht, beispielsweise inwiefern mögliche Ursachen für steigende oder fallende Bewertungen gefunden werden können. Dies ist auch wichtig, um eine Aussage bzgl. der Validität der kontinuierlichen Bewertungen zu treffen, etwa im Hinblick darauf, ob eine momentane oder akkumulierte Bewertung abgegeben wird. Zunächst werden allerdings die methodischen Vorge-

hensweisen bei der Analyse erläutert, um die spätere Datenanalyse übersichtlicher darstellen zu können.

5.2.1 Methodische Vorüberlegungen

Der erste Eindruck von einer deskriptiven Darstellung der Bewertungen der einzelnen Personen im Verlauf der vierten Performance (s. Abb. 5.7)¹, wobei jeder Graph die Echtzeit-Bewertung einer Person darstellt, ist „Chaos“. Die Zeitskala reicht vom Betreten der Bühne bis hin zum Verlassen, wobei in *emoTouch* der Zeiger am Anfang auf den Wert Null gesetzt wurde, was den einheitlichen Beginn der Bewertungen erklärt. Am Ende gibt es viele Sprünge an das untere Ende der Skala, da einige Versuchspersonen Probleme hatten, den *Exit*-Knopf zu treffen und stattdessen den Zeiger bewegt und somit eine plötzlich sehr schlechte Bewertung abgegeben haben. Ab diesen Sprüngen dürfen die Daten folglich nicht mehr ausgewertet werden, da dies für den jeweiligen Zuhörer dem Ende entsprach und auch niemand so frühzeitig die Bewertung beendet hat, dass nachträglich noch wichtige Teile der Performance unberücksichtigt blieben.

Ist nun lediglich die Darstellung unübersichtlich, oder sind die Bewertungen tatsächlich als inkonsistent anzusehen? Die Streuung der absoluten Wertungen ist anhand der Grafik beurteilt zu jedem Zeitpunkt hoch, mit Ausnahme der ersten wenigen Sekunden, auch wenn man berücksichtigt, dass das untere Ende der Skala, die von -100 bis +100 geht, abgeschnitten wurde. Die Verläufe sind anscheinend ebenfalls individuell, sodass wenn ein oder mehrere Graphen nach oben oder unten einknicken, die anderen dies zeitgleich meist nicht tun. Allerdings ist über den gesamten Zeitverlauf ein leichter Anstieg der meisten einzelnen Bewertungen zu erkennen.

Eine mögliche Erklärung für die sehr unterschiedlichen absoluten Bewertungen ist wiederum, dass die Probanden mit der Skala von „sehr schlecht“ bis „sehr gut“ unterschiedlich umgehen, also verschiedene Ansprüche stellen für eine gute Leistung oder manche mehr und andere weniger versuchen, die Skala komplett auszureizen. Ein Lösungsansatz dafür ist, die Bewertungen so zu standardisieren, dass Unterschiede in der Verwendung der Skala ausgeglichen werden. Dafür wurden die Bewertungen jeder einzelnen Person über den Zeitverlauf vom Beginn des ersten

¹Bei allen Graphiken, die Zeitreihen darstellen, werden aus Gründen der Übersichtlichkeit grundsätzlich Bewertungen einer Person oder auch Maße wie Mittelwert oder Quartile als Linie miteinander verbunden, auch wenn in den Daten tatsächlicher Sprung vorliegt, weil beispielsweise ein Proband zwischendurch den Finger weggenommen hat.

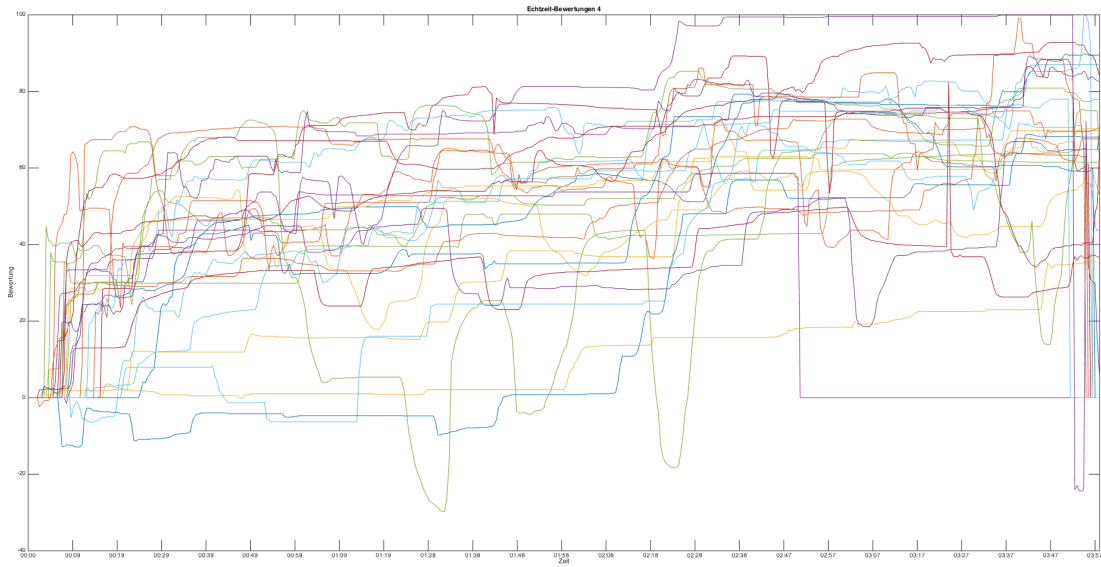


Abbildung 5.7: Die Bewertungen aller Probanden zu Performance 4

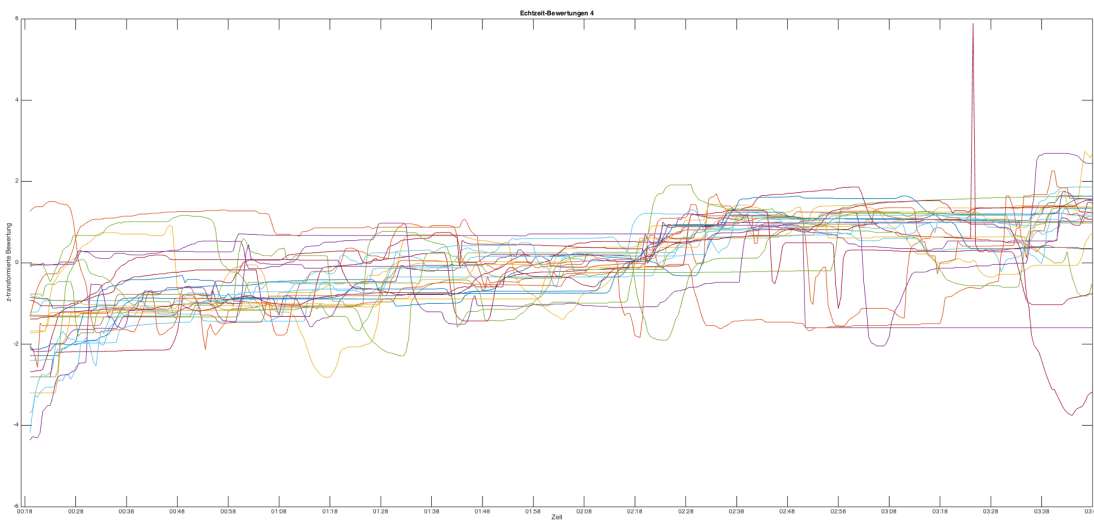


Abbildung 5.8: Die z-transformierten Bewertungen aller Probanden zu Performance 4

Tons bis zum Ende des letzten Tons z-transformiert, also die Bewertungen so entlang der Hoch-Achse verschoben, dass die mittlere Bewertung jedes Teilnehmers 0 ist und in gleicher Richtung gestreckt bzw. gestaucht, sodass die Standardabweichung jedes Teilnehmers 1 ist. Für die z-Transformation wurde die Zeit des Bühnenauftritts und -abgangs ausgeschlossen, da sonst die Validität insofern eingeschränkt würde, dass während des Bühnenauftritts ausschließlich außermusikalische Faktoren, also das Bühnenverhalten bewertet werden kann. Sobald Musik klingt ist hingegen, ist diese der zentrale Aspekt der Performance und das Bühnenverhalten weniger wichtig als vorher. Die Benutzung der Skala könnte sich beispielsweise unterscheiden, je nachdem, ob eine musikalische Performance in dem Moment bewertet wird, oder der Bühnenauftritt und lediglich die Erwartung auf die Musik mitschwingt, bzw. bei dem Abgang das zuletzt Gehörte noch nachklingt. Die so berechneten Bewertungsverläufe sind in Abb. 5.8 dargestellt, wo der optische Eindruck zunächst ist, dass die Graphen dichter verlaufen. Es scheint durch die Anhäufung vieler Graphen eine globale Kurve sichtbar zu werden, von der lediglich eine gewisse Menge an Ausreißern abweicht. Das Bild entsteht dadurch, dass Graphen, die vorher in den Originalbewertungen im Chaos parallel verliefen, nur mit Mühe zu erkennen waren, während sie hier offensichtlich auf fast gleicher Höhe verlaufen. Allerdings bleibt zu berücksichtigen, dass dieser Eindruck dadurch verstärkt wird, dass um die Ausreißer darstellen zu können, speziell den bei ca. 3:25 min, die Skala von -6 bis +6 aufgetragen ist, also je bis zur sechsfachen Standardabweichung über und unter dem arithmetischen Mittel. Dieser kurze sprunghafte Wechsel zu einer bedeutend besseren Bewertung und sofort zurück wird erst durch die z-Transformation so deutlich sichtbar, da er von den absoluten Werten her im mittleren Bereich aller Bewertungen ist, aber man kann ihn auch bereits in den Originalbewertungen sehen. Wahrscheinlich ist der Sprung darauf zurückzuführen, dass der Proband versehentlich abgerutscht ist oder mit einem anderen Finger eine Stelle weiter rechts auf dem iPad-Display berührt hat, und im nächsten Moment diese unabsichtlich andere Bewertung wieder zurück korrigiert hat. Folglich ist dem keine große inhaltliche Bedeutung zuzumessen. Die z-Transformation macht insofern die Darstellung übersichtlicher, sodass plötzliche starke Bewertungsänderungen besser auffallen und ähnliche Bewertungsverläufe besser zu erkennen sind. Allerdings können durch diese Art der Standardisierung auch inhaltliche Fehler passieren. Beispielsweise könnte ein Proband bewusst den Zeiger nur minimal bewegt haben, weil für ihn die Leistung durchgehend fast gleich gut war. Durch die z-Transformation würden dann die kleinen Bewegungen verstärkt und anschließend vermutlich als inhaltliche Unterschiede interpretiert.

Speziell einzelne Ausreißer, auch die mit inhaltlicher Begründung und ggf. sogar recht kleiner Amplitude, bei sonst sehr gleichbleibender Bewertung, würde durch die z-Transformation ein riesiger Bewertungsunterschied. Dieser Fall trifft beispielsweise bei einer kurzzeitig bedeutend schlechteren Bewertung, weil ein Musiker sich verspielt hat oder gar herausgeflogen ist und an einer Stelle neu anfangen muss, zu, wenn ansonsten wenig Unterschiede in der Leistung angegeben werden. Folglich sollten kurzzeitige deutliche Bewertungsänderungen einzelner Personen in der z-transformierten Darstellung mit Vorsicht ausgewertet werden. Auf der anderen Seite können auch als sehr groß empfundene Leistungsunterschiede und entsprechend große Bewertungsunterschiede im Zeitverlauf durch die Transformation deutlich verringert werden, sodass sie möglicherweise auf einer Stufe mit den bewusst kleinen Veränderung eines anderen Probanden stehen. Ein weiteres Problem dabei ist, dass alle Zuhörer zu einem einheitlichen Anfangspunkt gezwungen wurden, also ihre persönliche Verwendung der Skala erst später zum Tragen kommt. Durch die Einschränkung auf die reine Spielzeit der Musikstücke, sind die Anfangsbewertungen zwar nicht mehr zwingend bei 0, allerdings oft noch in der Nähe und somit für dieses und auch die meisten anderen Stücke der Tiefpunkt vieler Bewertungsverläufe - der sich allerdings auf Mittelwert und Standardabweichung auswirkt und somit auf die z-Transformation. Verschieden hohe Bewertungen und verschieden starke Bewertungsschwankungen aneinander anzupassen, ist gleichzeitig die Stärke und Schwäche der z-Transformation und in diesem Kontext führt sie zu einer deutlichen Verbesserung der Vergleichbarkeit der Verläufe und wird insofern verwendet. Für die Konsistenz der absoluten Bewertungen kann allerdings mit z-transformierten Daten keine Aussage mehr getroffen werden.

Um die Datenmenge für die Auswertung zu reduzieren, wurden zunächst Mittelwert und Standardabweichung berechnet, dargestellt in Abb. 5.9 und 5.10 jeweils als Mittelwert \pm Standardabweichung basierend auf den ursprünglichen bzw. den z-transformierten Daten. Erst dadurch sind die Abweichungen zu einzelnen Zeitpunkten messbar und die Analyse des durchschnittlichen Bewertungsverlauf wird möglich. Für die Interpretation dieser Kennwerte kann allerdings es wichtig sein, zu wissen, wie die Daten verteilt sind. Die bereits anhand aller Daten beobachtete über die Zeit ansteigende Tendenz der Bewertungen ist nun deutlich zu sehen. Speziell am Anfang während des Bühnenauftritts bis 19s und auch noch kurz danach ist ein starker Anstieg erkennbar bis dieser bei ca. 40s mit einer mittleren Wertung von 40 abflacht und bis ca. 2:20 min die Bewertungen nur noch minimal ansteigen. Dann ist nochmals ein deutlicher Anstieg zu verzeichnen, ein leichter Abfall bei

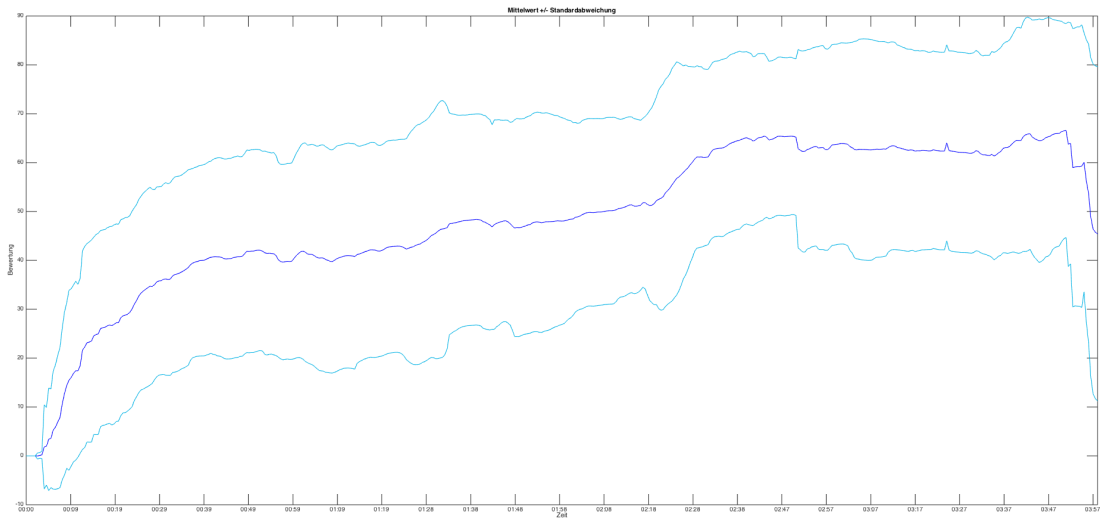


Abbildung 5.9: Mittelwert \pm Standardabweichung aller Bewertungen zu Performance 4

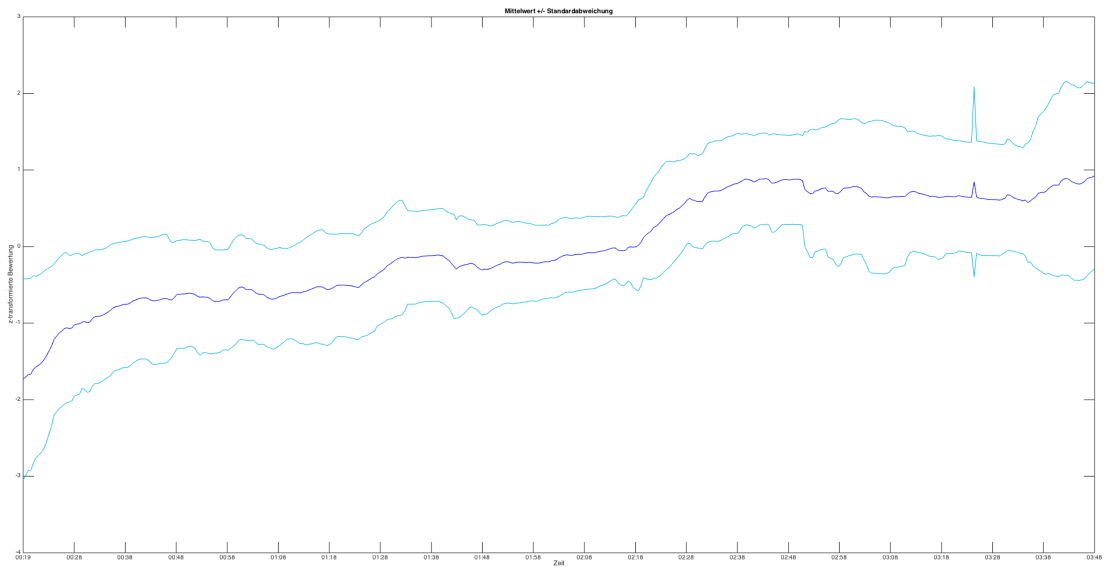


Abbildung 5.10: Mittelwert \pm Standardabweichung aller z-transformierten Bewertungen zu Performance 4

knapp 3 min und danach bleibt das arithmetische Mittel bis kurz vor dem Ende, wo es nochmals leicht ansteigt, konstant bei ca. 60. Die Abfälle ganz am rechten Rand der Grafik werden wegen der bereits genannten technischen Probleme nicht weiter berücksichtigt. Man kann also schließen, dass die Performance entweder über die Zeit immer besser wurde oder evtl. auch konstant gut war und die Zuhörer durch Akkumulation der Eindrücke zu immer besseren Bewertungen kamen. Speziell bzgl. des Anstiegs bei etwa 2:30 min ist zu vermuten, dass die Musikerin das Publikum besonders beeindruckt hat und bei dem leichten Abfall kurz darauf möglicherweise ein Fehler passiert ist, was nun mit der Videoaufzeichnung abgeglichen werden könnte. Im Abschnitt 5.2.3 auf Seite 82 wird exemplarisch für ein der Stücke versucht, Bezüge zwischen den Echtzeitbewertungen und der Performance herzustellen, aber an dieser Stelle sei nur auf diese Möglichkeit hingewiesen, da weiterhin die methodischen Fragen mit dem Ziel einer Aussage über Konsistenz verfolgt werden. Über die Veränderung der Standardabweichung lässt sich anhand von Abb. 5.9 nicht viel aussagen, außer dass sie nahezu konstant bleibt, da die Graphen für Mittelwert plus bzw. minus eine Standardabweichung quasi parallel zur Mittelwertkurve verlaufen. Lediglich bei 2:00 bis 2:20 min und ganz am Anfang beim Bühnenauftritt ist sie etwas geringer.

Bei der gleichen Analyse anhand des Mittelwerts und der Standardabweichung der z-transformierten Daten in Abb. 5.10 scheidet eine Belastung der absoluten Mittelwerte aus, genauso wie die Analyse des Bühnenauftritts und Abgangs. Davon abgesehen, ist der Verlauf aber vergleichbar dargestellt und die Standardabweichung verändert sich stärker, sodass hier viel eindeutiger zu erkennen ist, dass im Zeitfenster von 2:00 bis 2:20 min eine bessere Einigkeit über die Leistung besteht. Außerdem fällt auf, dass die Streuung zwischen 3:00 und 3:18 min nochmals größer ist und zu Beginn des Stückes die Streuung größer statt kleiner, wie bei den Originaldaten, ist. Den Anfang sollte man dementsprechend, wie bereits erwähnt, in der z-Transformation mit Vorsicht betrachten, während über den weiteren Verlauf die Streuung der transformierten Daten als aussagekräftiger angesehen werden kann. Ein wesentliche Vorteil ist, dass parallel verlaufende Bewertungskurven hier nicht mehr die Standardabweichung künstlich erhöhen, ohne dass dies mit einer tatsächlichen Uneinigkeit über die Leistung gleichzusetzen ist. Der Ausreißer bei ca. 3:25 min ist bei beiden Varianten, aber besonders verstärkt bei der standardisierten Daten in die Mittelwertdarstellung, eingegangen, und verzerrt hier sogar die Standardabweichung. Dieses Problem ist bei der Betrachtung dieser beiden deskriptiven Maße nicht auszuschließen und eine Verzerrung durch Ausreißer, speziell durch

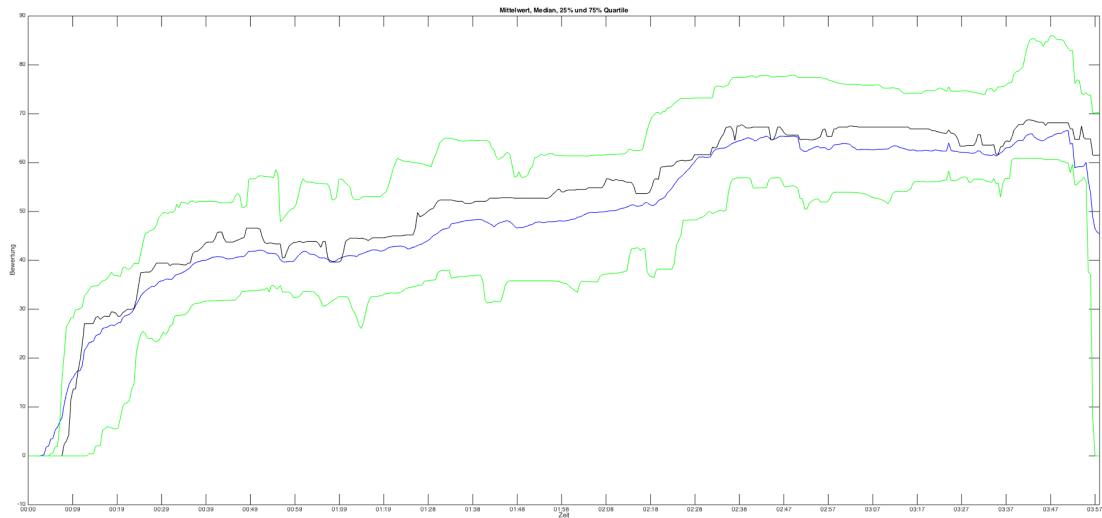


Abbildung 5.11: Mittelwert, Median und Quartile der Bewertungen zu Performance 4

jene, die nicht an plötzlichen Peaks erkennbar sind, ist an noch weiteren Stellen denkbar.

Um eine Verzerrung durch Ausreißer und eine Schiefe der Verteilung aufzudecken, hilft ein Vergleich des Mittelwerts mit dem Median, ohne für jeden Zeitpunkt eine genauere Analyse der Verteilung zu machen, die schwierig übersichtlich darzustellen ist. In Abb. 5.11 ist zu erkennen, dass der Median fast durchgängig höher als die Standardabweichung ist, insofern also mehr als 50 % der Bewertungen oberhalb des Mittelwerts liegen. Vermutlich sind viele nur knapp höher, während der Mittelwert durch wenige Ausreißer nach unten deutlich verringert wird. Dieses Phänomen wird durch die z-Transformation (s. Abb. 5.12) teilweise ausgeglichen, insofern dass der Median nicht durchgehend höher ist, sondern nur noch in bestimmten Abschnitten eine beachtenswerte Differenz zwischen Mittelwert und Median besteht. Der Median hat bei der Betrachtungen von Zeitreihen allerdings den Nachteil, dass er immer der Bewertung einer Person entspricht, da die Daten einer ungeraden Zahl von Probanden ausgewertet werden. Infolgedessen spiegelt der Verlauf des Medians nicht die Bewertungen insgesamt wieder, sondern lediglich den einer Person, bis dieser sich mit dem Bewertungsverlauf einer anderen Person kreuzt und dann die Bewertung dieser Person für den nächsten Zeitabschnitt der Median ist. Gleiches gilt auch für die Quartile, also die Werte, für die 25 bzw. 75 % der Bewertungen darunter oder darüber liegen. Sie können allerdings als alternatives Streuungsmaß



Abbildung 5.12: Mittelwert, Median und Quartile der z-transformierten Bewertungen zu Performance 4

zur Standardabweichung benutzt werden, die den Nachteil hat, die Richtung der Abweichung vom Mittelwert nicht zu berücksichtigen. Die Standardabweichung zu einem bestimmten Zeitpunkt ist ein Wert, der in Abb. 5.9 und 5.10 sowohl zum Mittelwert addiert als auch davon subtrahiert dargestellt wurde, unabhängig davon, wie stark die Werte tatsächlich in die eine oder andere Richtung streuen. Eine schiefe Verteilung beispielsweise ist nicht nur im Vergleich von Mittelwert und Median, sondern auch durch Median und Quartile gut zu erkennen. Der Abstand zwischen dem oberen Quartil und dem Median ist für den gesamten Verlauf, abgesehen von etwa der letzten halben Minute, deutlich geringer als der zwischen dem Median und dem unteren Quartil, sodass viel mehr Bewertungen nur knapp höher sind als der Median. Hingegen sind die 25% der Bewertungen zwischen Median und unterem Quartil deutlich breiter gestreut. Eine weitere Aussage kann durch den Vergleich von Mittelwert abzüglich Standardabweichung bzw. dem unteren Quartil für den Zeitraum des Bühnenauftritts, also den ersten 19 s bei dieser Performance, getroffen werden: Die Standardabweichung vermittelt den falschen Eindruck, die Bewertungen würden nach Beginn zunächst genauso nach unten, wie nach oben streuen (s. Abb. 5.9). Das untere Quartil hingegen bleibt bei 0 statt zu fallen und steigt nach einer etwas längeren Verzögerung dann ebenfalls an. Dies lässt eher auf eine große Anzahl unentschlossener Probanden schließen, die sich schwer tun, den

Bühnenauftritt mit einer so konkreten Bewertung zu versehen oder einfach länger brauchen um einen ersten Eindruck zu gewinnen und diesen auch in Form einer Verschiebung des Zeigers angeben. Außerdem könnte der erste Eindruck natürlich auch neutral sein, allerdings sollte es bei der feinen Unterteilung der Skala auf dem iPad, bei dem zunächst 1024 verschiedene Punkte erkannt werden, schwierig sein, den Finger aufzusetzen ohne den Zeiger zu verschieben. Um diesen Eindruck, dass (fast) niemand negativ bewertet, zu bestätigen, wurde das Minimum, welches im Zeitraum des Bühnenauftritts leicht negativ wird, und das 10% Quantil benutzt, welches entsprechend dem unteren Quartil so lange 0 ist, bis es in den positiven Bereich ansteigt. Dies ist auch in Abb. 5.7 auf Seite 73 daran sichtbar, dass lediglich zwei Personen innerhalb der ersten 19s leicht negative Bewertungen abgeben.

Nachdem nun das methodische Repertoire zur Analyse der Zeitreihen detailliert vorgestellt wurde, muss noch erwähnt werden, dass aus Gründen der Anschaulichkeit ein Stück mit sehr simplem mittleren Bewertungsverlauf ausgewählt wurde, ohne viele besonders auffällige Stellen und ohne nennenswerte Schwankungen der Standardabweichung. Ein weiterer Grund für die Auswahl war, dass bis zu diesem Stück alle Probanden im Umgang mit dem Interface sicher waren und es keine Frühstarts mehr gab, also niemand vorzeitig mit der Bewertung begonnen hat. Folglich ist eine Analyse des Videos anhand dieser Daten weniger interessant, als für einige andere Stücke, die im Abschnitt 5.2.3 auf Seite 82 noch genauer betrachtet werden.

5.2.2 Innere Konsistenz der kontinuierlichen Bewertungen

Eine Beurteilung, wie konsistent die Ergebnisse im Zeitverlauf sind, ist trotz aller methodischer Überlegungen schwierig. Zunächst wird die interindividuelle Übereinstimmung als Mittelwert der Standardabweichung zu jedem einzelnen Zeitpunkt der Performance inkl. Bühnenauf- und -abgang betrachtet, wobei die einzelnen Werte Tabelle 5.3 zu entnehmen sind. Die Streuung ist demnach bei Ensemble zwei und drei, dicht gefolgt von Ensemble 1 am höchsten, während sie bei den letzten beiden etwas geringer ausfällt, was auch bzgl. der Rangordnung der Streuungshöhe mit der für die Gesamtwertungen übereinstimmt (vgl. Abschnitt 5.1.1). Über alle Performances nochmals gemittelt ergibt sich eine Standardabweichung von 24.69, was unter Annahme einer Normalverteilung bedeutet, dass 68% der Bewertungen innerhalb eines Viertels der Skala sind. Ob dies eine hohe oder niedrige Konsistenz ist, lässt sich allerdings schwierig beantworten, da auch hier von den absoluten Bewertungen ausgegangen wird. Das ist aufgrund der uneinheitlichen Nutzung

der Skala ungünstig (vgl. Abschnitte 5.1 und 5.2.1) Bei den mittleren Standardabweichungen der für jeden Teilnehmer z-transformierten Zeitreihen ändert sich die Rangfolge teilweise, bei welcher Performance die Streuung größer oder geringer ist. Allerdings sind alle Werte kleiner als Eins, woraus folgt, dass die interindividuelle Streuung geringer ist als die Bewertungsänderung jeder einzelnen Person über die Zeit, da die Standardabweichung dafür durch die z-Transformation 1 ist. Auffällig ist, dass die Schwankungen der Standardabweichung über die Zeit im Vergleich zur durchschnittlichen momentanen Streuung, jeweils bezogen auf die nicht-z-transformierten Daten, gering sind. Folglich ist zu erwarten, dass eine ähnlich große Streuung zu allen Zeitpunkten vorliegt und entweder Stellen mit bedeutsam größerer oder kleiner Uneinigkeit der Hörer nur von kurzer Dauer oder die Schwankungen der Standardabweichung nur marginal sein können.

Schubert (2010, 234f.) schlägt vor, aus dem Mittelwert der Standardabweichungen zu den einzelnen Zeitpunkten \bar{X} und der Standardabweichung dieser Standardabweichungen $\bar{\sigma}$, im Folgenden analog zu Schubert als Standardabweichung zweiter Ordnung bezeichnet, eine Schwelle τ zu berechnen, anhand der die Ratings zu einzelnen Zeitpunkten als signifikant bzw. reliabel eingestuft werden können. Die Schwelle berechnet sich aus der mittleren Standardabweichung abzüglich eines (kleinen) Vielfachen k der Standardabweichung zweiter Ordnung. Als Formel ausgedrückt ergibt sich $\tau_{-k} = \bar{X} - k\bar{\sigma}$, wobei die Variablennamen überwiegend von Schubert (2010, 234f.) übernommen sind. Lediglich τ_{-k} für die Schwelle wird analog zu Schubert (2013) verwendet. Schubert (2013) benutzt in der Formel alternativ \pm , sodass die Schwelle auch größer als die mittlere Standardabweichung sein kann, wobei bei Schubert (2010) k als kleine Zahl definiert ist. Vermutlich ist damit eine kleine positive Zahl gemeint ist, aber es wird nicht explizit ausgeschlossen, dass k negativ sein könnte, was dann der späteren Variante der Formel entsprechen würde. Die Schwelle τ_{-k} ist nun als obere Schranke der Standardabweichung zu verstehen, bis zu der die Bewertungen zu einem bestimmten Zeitpunkt als reliabel angesehen werden können (Schubert, 2010, 2013). Aus der Berechnung ergibt sich, dass dies eine Abschätzung darstellt, wann innerhalb des gesamten Zeitraums die Streuung der Bewertung größer oder geringer ist. Folglich kann so keine allgemeingültige Reliabilität festgestellt werden, wie Schubert in beiden Publikationen auch selbst argumentiert.

Diese Schwelle könnte angewandt werden um Stellen mit deutlich geringerer interindividueller Streuung zu finden, indem das k betragsmäßig größer als 1 oder 0.5, wie von Schubert vorgeschlagen, gewählt wird. Somit würden nicht große Teile des

Stück	1	2	3	4	5	Mittelwert gesamt
mittlere Standardabweichung im Zeitverlauf	26.36	29.27	29.13	20.49	18.19	24.69
Standardabweichung der Standardabweichung im Zeitverlauf	3.10	4.62	5.19	3.85	6.15	4.58
mittlere Standardabweichung der z-transformierten Bewertungen im Zeitverlauf	.896	.888	.805	.735	.794	.823

Tabelle 5.3: Standardabweichung der Zeitreihen

Stücks unterhalb der Schranke liegen, sondern nur einige wenige interessante Stellen. Man könnte das Kriterium auch andersherum anwenden, also mit $k \leq -1$, um Stellen mit besonders großer Streuung ausfindig zu machen. Diese Schwankungen können nicht übergreifend über alle Daten betrachtet, sondern müssen gezielt anhand der Zeitreihen analysiert werden. Außerdem bezieht sich diese Frage auch bereits auf die Konsistenz der Bewertungsverläufe, und wird somit im folgenden Abschnitt genauer betrachtet.

Als Fazit dieses Abschnitts werden allerdings noch zwei der Forschungsfragen bzgl. der Analyse der Zeitreihen beantwortet. Zu Frage 2a, wie stark die Standardabweichung schwanke, lässt sich festhalten, dass die Standardabweichung zweiter Ordnung sowohl im Hinblick auf die Skala als auch auf die momentane Streuung der Bewertungen sehr gering ist. Entsprechend verändert sich die Standardabweichung im Verlauf der Stücke nur wenig. Zur Reliabilitätsfrage 2b lässt sich allerdings so allgemein nur schwierig etwas sagen, da Vergleichswerte oder ein sinnvolles Kriterium, das angewendet werden kann, fehlen. Anhand der Zeitreihen können möglicherweise Unterschiede festgestellt werden, sodass die Reliabilität im Vergleich zu bestimmten Zeitpunkten besser oder schlechter ist. Allerdings lässt sich auch dort nur innerhalb der Stücke bzw. zwischen den Stücken vergleichen und keine objektive Aussage treffen. Die Frage bleibt insofern offen.

5.2.3 Konsistenz der Bewertungsverläufe und Performanceanalyse mit Hilfe der kontinuierlichen Bewertungen

In diesem Abschnitt wird anhand von einem exemplarisch ausgewählten Stück der Bewertungsverlauf und dessen Konsistenz ausführlich untersucht. Bzgl. der übrigen Stücke werden nur einzelne Beobachtungen, die sich von dem Beispielstück

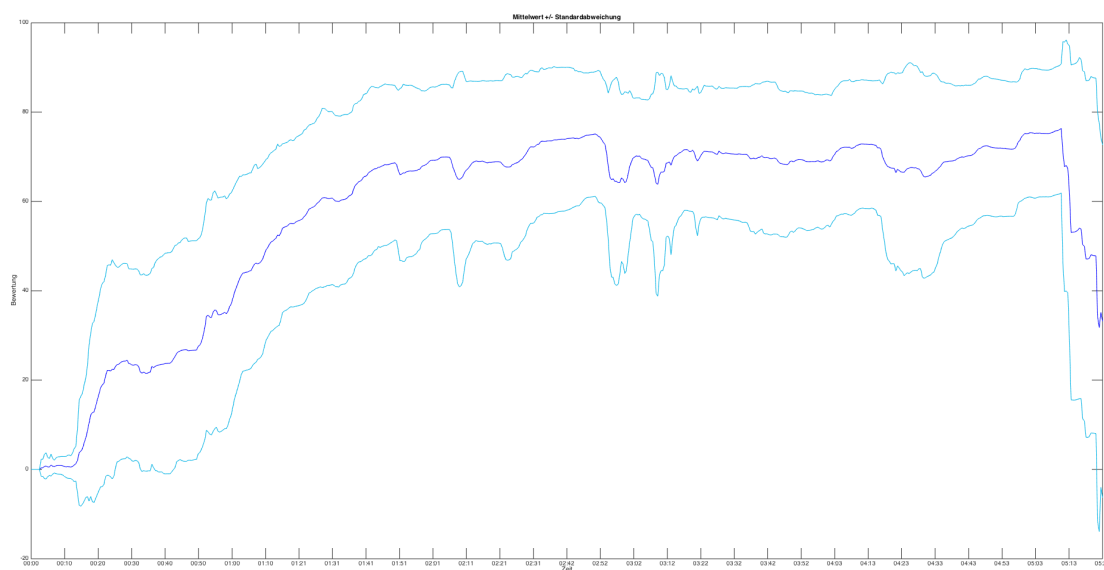


Abbildung 5.13: Mittlere Bewertung \pm Standardabweichung von Performance fünf im Zeitverlauf

unterscheiden, herausgegriffen, und daher nicht unberücksichtigt bleiben können. Eine vollständige Analyse wäre im Rahmen dieser Arbeit nicht möglich. Dabei wird versucht, die Erkenntnisse direkt mit der Video-Aufzeichnung der Auftritte in Verbindung zu bringen und zu analysieren, worauf die Zuhörer vermutlich reagiert haben, wenn sie Ihre Bewertung verändert haben.

Performance 5

Die Darbietung des fünften Ensembles am Vorspielabend wurde für die detailliertere Analyse ausgewählt, weil global ein simpler Bewertungsverlauf vorliegt, aber gleichzeitig auch einige kurzzeitige Schwankungen auftreten, die in Bezug zu dem Video betrachtet werden können. Das Ensemble spielte das Stück „Tall Fiddler“ von Tommy Emmanuel, welches dieser laut Aussage in seinem eigenen YouTube-Clip mit der Absicht, einen Geiger nachzuahmen, geschrieben hat (Emmanuel, 2013). Um die Konsistenz der Bewertungsverläufe zu untersuchen ist, wie bereits erläutert, die z-Transformation die sinnvollere Darstellungsform. Allerdings werden zunächst kurz Mittelwert und Standardabweichung der Originaldaten betrachtet, um den Bühnenaufgang und auch anschließend den Abgang miteinbeziehen zu können. Der globale Verlauf der mittleren Bewertung ist ein recht steiler Anstieg zu Beginn



Abbildung 5.14: Mittlere z-transformierte Bewertung \pm Standardabweichung von Performance fünf im Zeitverlauf

von einer Dauer von knapp zwei Minuten und anschließend eine gleichbleibende nur wenig schwankende Bewertung. Diese ist mit 70 nah an dem oberen Ende der Skala. Dabei ist die Standardabweichung sowohl beim Anstieg als auch später weitgehend konstant, sodass davon ausgegangen werden kann, dass dieser globale Verlauf von allen Zuhörern gewissermaßen ähnlich vollzogen wurde. Es ist keine merkliche Reaktion auf den Beginn des Stücks nach 55.15s oder das Ende bei 5:09 min zu erkennen, abgesehen davon, dass ab 5:12 min die mittlere Bewertung nach und nach sprunghaft fällt. Diese Reaktion passiert deutlich bevor der Applaus endet und die Musiker die Bühne verlassen haben, somit also für ein Beenden der Session sehr früh, speziell unter der Berücksichtigung einer gewissen Latenz um überhaupt auf das Ende des Stückes zu reagieren. Dennoch kann hier nicht sicher von einer schlechteren Bewertung ausgegangen werden, sondern davon, dass bereits hier versucht wird, den *Exit*-Knopf zu drücken, und dabei versehentlich die Bewertung heruntermgesetzt wird. Eine Auswertung des Bühnenabgangs ist somit kaum möglich. Bei dem Bühnenauftritt, der bei diesem Ensemble mit ca. 55s gut doppelt so lange dauerte wie bei allen anderen, fällt auf, dass wieder der Mittelwert abzüglich der Standardabweichung kurz nach Beginn nach unten in den negativen Skalenbereich ausschlägt. Allerdings lässt sich, wie in Abschnitt 5.2.1 für die vierte Performance gezeigt, widerlegen, dass eine nennenswerte Anzahl an Probanden

negativ bewertet hat, da beispielsweise das untere Quartil nicht kleiner als Null wird. Des Weiteren gibt es in dem steilen, bis deutlich nach Beginn des Stücks anhaltenden Anstiegs einen kleinen Einschnitt bei 30 bis 45 s, der diesen einen Moment verzögert. Diese Publikumsreaktion ist am ehesten auf die Verzögerung des Aufgangs, möglicherweise auf das doppelte Einstellen des Notenständers, zurückzuführen. Der Gitarrist fragt zwar vor Beginn noch, ob inzwischen alle „fertig mit Start drücken“ (ca. 0:45-0:50 min) sind, allerdings steigt ab diesem Punkt die Bewertung wieder und die Nachfrage wurde zwar mit Lachen kommentiert, aber offensichtlich durchaus positiv aufgenommen.

Die Analyse der Spielzeit wird mit Ausnahme des Anfangs, da hier durch den vereinheitlichten Startwert noch eine Verzerrung vorliegt, anhand der z-transformierten Daten analysiert. Die Darstellung des Mittelwerts \pm Standardabweichung in Abb. 5.14 zeigt, dass hier die ohne Normierung kleinen Mittelwertschwankungen deutlich verstärkt wurden und zudem auch Veränderungen der Standardabweichung sichtbar sind. Um ein Stück weit sicher zu stellen, dass nicht einzelne Ausreißer die Abfälle der mittleren Bewertung verursachen, wird zusätzlich mit dem Median verglichen. Im Verlauf lassen sich drei Stellen identifizieren, an denen die Standardabweichung merklich größer wird und die mittlere Bewertung mindestens leicht abfällt: Die erste Stelle von ca. 2:07 bis 2:30 min ist dabei die unauffälligste mit zwei kurzen, schwachen Rückgängen des Mittelwerts, bei gleichzeitigem Anstieg der Standardabweichung. Diese sind im Median aber nicht vorhanden. Das zweite interessante Zeitfenster von 2:50 bis 3:18 min zeichnet sich durch zwei kurze, aufeinander folgende und deutlich stärkere Ausschläge des Mittelwerts nach unten aus. Die Standardabweichung steigt gleichzeitig an, nachdem sie kurz vorher und kurz nachher deutlich geringer als im Mittel des Verlaufs ist. Der Median hat hier nur bei dem ersten Mittelwertabfall ebenfalls ein lokales Minimum, allerdings nimmt beim zweiten Ausschlag auch das untere Quartil einen deutlich niedrigeren Wert an. Obwohl die obere Hälfte der Bewertungsverläufe die Veränderung nicht mitmacht, scheinen sich die kritischeren Hörer mit Bewertungen unterhalb des Medians darüber einig zu sein, dass diese Stelle schlechter performt war. Die beschriebenen Stellen sind visuell so gut zu erkennen, dass die Reliabilitätsschwelle von Schubert keinen Vorteil bringen würde bzw. sogar eher einen Nachteil, da nur ein bestimmter Grenzwert gesetzt wird. Anhand der graphischen Darstellung ist eine exakte Schwelle zwar schwierig zu erfassen, aber dafür kann die Entwicklung deutlich besser berücksichtigt werden. Aufgrund dessen wird das Kriterium hier nicht angewandt.

Im Video sind in diesen Zeitfenstern, die nach vorne um einige Sekunden erweitert wurden, um die Reaktionszeit zu berücksichtigen, diverse mögliche Ursachen für die Bewertungsänderungen erkennbar. Die erste Stelle wird dabei wegen fehlender Veränderung des Medians nicht betrachtet und die Video-Analyse beginnt mit der zweiten genannten Stelle. Bei 2:48 min setzt die Geige ab, nachdem sie gerade die ersten paar Töne des Themas gespielt hatte. Auf dem Video wird es nur wenig deutlich, im Vergleich zu Live-Situation, in der ich als ZuhörerIn zumindest sicher war, dass sie an der Stelle gar nicht hätte spielen sollen. Im Video kann man dies eher an der Reaktion des Gitarristen erkennen. Kurz danach ab ca. 2:51 wird die Bewertung schlechter und für einige Takte spielen nur der Gitarrist und der Bassist. Zunächst spielt der Bassist ein Solo mit Begleitung der Gitarre, wobei das Publikum möglicherweise die Intonation zu schlecht fand. Bei 3:04-3:06 min ruft der Gitarrist noch etwas zu seinen Mitspielern, speziell in Richtung der Violinistin, und direkt danach geht das Stück wieder mit allen mit dem Hauptthema weiter. Es folgt allerdings erst in diesem Moment, wo das Stück wieder deutlich sicherer läuft und mehr mitreißt, der Abfall der Bewertung, die dann erst entsprechend verzögert wieder besser wird. Es ist an dieser Stelle schwierig zu sagen, wie groß die Latenzzeit ist und worauf das Publikum mit der schlechteren Bewertung reagiert hat. Für den weiteren Bewertungsverlauf ab dieser Stelle sind zwei verschiedene Interpretation möglich, nämlich zum Einen, dass die kurzen Aussetzer die Bewertung nicht längerfristig beeinflussen, da der Mittelwert sofort wieder leicht steigt und bis zum Ende des Stücks auf einem ähnlichen Level bleibt. Direkt vor dem Rückgang bei 2:50 min ist noch ein weiterer Anstieg mit sinkender Standardabweichung zu erkennen, also mit hoher Einigkeit darüber, dass es tendenziell noch besser wird. Eventuell wurde dieser Anstieg durch die Verspieler und den weniger mitreißenden solistischen Zwischenteil aufgehalten und dadurch anschließend nicht fortgeführt wird. Folglich stagniert die Bewertung wegen dem als schwächer wahrgenommenen Abschnitt bei ca. 70 und hätte sich möglicherweise andernfalls noch höher einpendeln können.

Die dritte, anhand der Zeitreihen identifizierte Stelle von 4:12 bis 5:00 min zeichnet sich wie die vorhergehenden durch einen Rückgang des Mittelwerts aus bei Anstieg der Standardabweichung, der hier wieder im Median nicht zu erkennen ist. Dabei ist der kurzzeitige Anstieg des Medians so fein, dass er nicht ausgewertet wird, da er zu stark durch den Bewertungsverlauf einer einzelnen Person beeinflusst sein kann. Die Bewertung wird im Bezug auf das Video ab dem Zeitpunkt schlechter, als der Gitarrist nochmals, diesmal für eine etwas längere Passage, bis auf kurze Einwürfe

komplett alleine spielt und sich dabei am Anfang auch verhaspelt mit den schnellen Läufen. Erst verzögert wird die Bewertung wieder besser. Eine nähere Analyse der möglichen weiteren Ursachen findet hier nicht statt, da das Zeitfenster so groß ist, dass es noch schwieriger ist, als bei der letzten Stelle, im Video eindeutige Bezüge zu finden. Zusammenfassend lässt sich feststellen, dass Passagen, in denen nicht alle Beteiligten spielten und das Folkstück einen Teil seines mitreißenden Charakters verlor, bei den Zuhörern im Vergleich schlechter ankamen.

Weitere Performances und Schlussfolgerungen

Für Ensemble zwei zeigt sich ein deutlich abweichender globaler Verlauf gegenüber den bisher betrachteten Stücken 4 und 5 (vgl. Abb. 5.15). Zunächst steigt die mittlere Bewertung während des Bühnenauftritts der Musikerinnen leicht an, macht einen kleinen aber deutlichen Sprung nach oben nachdem bei 32s der erste Tone erklingt. Danach flacht sie allerdings ab und bleibt konstant, solange nur die Pianistin die Einleitung spielt. Nach ca. 1:18 min setzen die Sängerinnen ein und ca. zwei Sekunden später sinken die Bewertungen im Mittel deutlich ab. Über einen recht langen Zeitraum setzt sich dann ein leicht fallender Trend bei gleichzeitig leicht ansteigender Standardabweichung fort. Ab 3:04 min steigt die Bewertung wieder und bleibt bis zum Schluss ansteigend, aber an dieser Stelle ließ sich kein eindeutiger Einflussfaktor finden – speziell nicht dafür, dass die Bewertungen danach stetig besser werden. Bei diesem Stück verändert sich auch die Standardabweichung der einzelnen Bewertungen über die Zeit merklich, allerdings über einen so langen Zeitraum, dass es nicht möglich ist, einen bestimmten Auslöser anhand des Videos dafür zu finden. Auch hier wäre es nicht zielführend, die Reliabilitätsschwelle von Schubert (2010) anzuwenden, da dadurch ein bestimmter Zeitpunkt, wo sie überschritten wird, gefunden werden könnte, aber dieser bei einer langsamen Entwicklung nicht aussagekräftig ist. Auch für Performance Nr. 3 ergibt sich ein deutlich anderer Verlauf, der zwar im Mittel bei Null beginnt, aber dann nicht steigt und ab einem gewissen Punkt konstant bleibt. Stattdessen verändert sich die mittlere Bewertung bis zum Schluss und beinhaltet deutlich schlechtere Bewertungen, bei allerdings nahezu konstanter Standardabweichung. Bei diesem Duett sind viele offensichtliche Fehler passiert bzw. die hohen Töne klangen häufig schlecht und außerdem waren die beiden Musikerinnen merklich nervös während ihres Auftritts. Teilweise lassen sich die Fehler mit den Wertungsabfällen in Verbindung bringen, allerdings reagiert das Publikum im Mittel nicht immer, wenn sich die Duettpartnerinnen verloren bzw. wieder ein Abschnitt kam, der besser gespielt wurde.

5 Ergebnisse

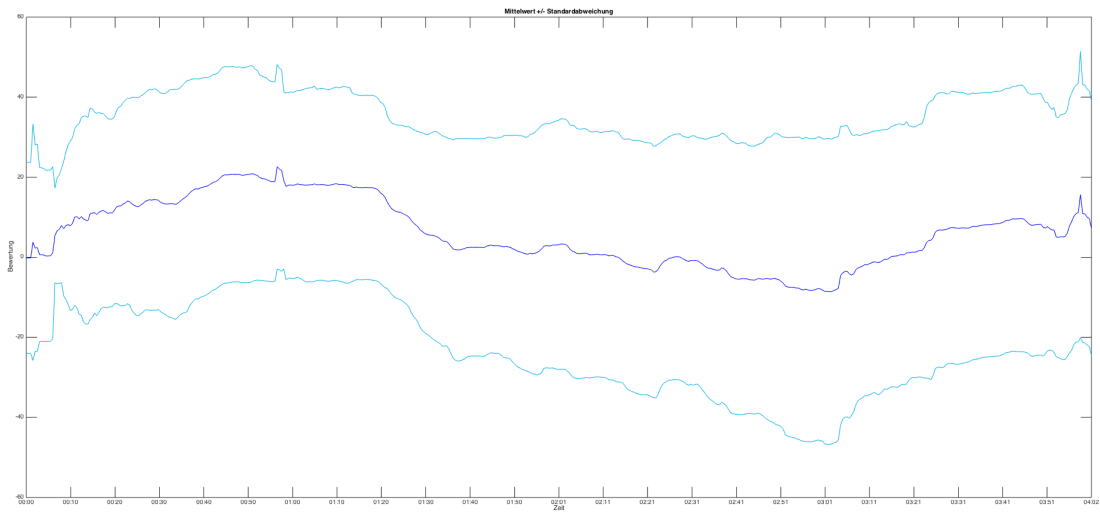


Abbildung 5.15: Mittlere Bewertung \pm Standardabweichung von Performance zwei im Zeitverlauf

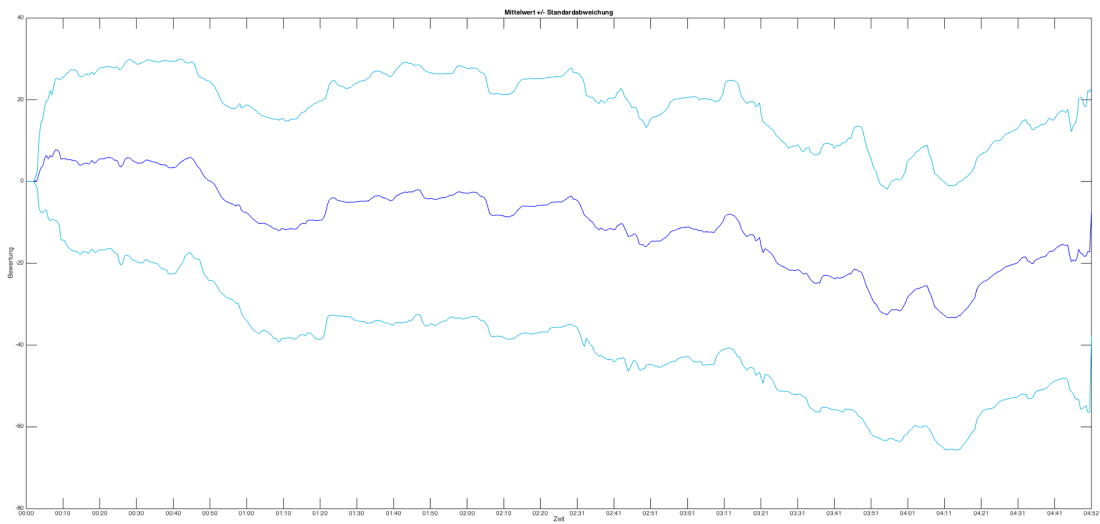


Abbildung 5.16: Mittlere Bewertung \pm Standardabweichung von Performance drei im Zeitverlauf

Die noch offenen Forschungsfragen bzgl. des Bewertungsverlaufs lassen sich mit diesen Erkenntnissen immerhin teilweise beantworten. Frage 2c, wann im Verlauf die Übereinstimmung der Bewertungen höher oder niedriger sei, hat keine große Relevanz, da sich die Standardabweichung im Zeitverlauf nur sehr wenig verändert. Einzelne Zusammenhänge zwischen kurzfristigen Bewertungsänderungen und plausiblen Ursachen lassen sich zwar herstellen, aber überwiegend passieren die Veränderungen über längere Zeiträume, sodass dies nicht möglich ist. Insofern ist Forschungsfrage 2e nach der Art des Zusammenhangs mit den Aufnahmen ähnlich zu beantworten. Es geht stellenweise, plausible Zusammenhänge zu finden, aber lange nicht immer. Die Erkenntnisse sind weit davon entfernt, erklären zu können, auf welche Events in einer Performance die Zuhörer mit einer Bewertungsänderung reagieren und auf welche eher nicht. Forschungsfrage 2d, ob es gleichförmige Verläufe gebe, kann weitestgehend bejaht werden, insofern dass deutliche Mittelwertschwankungen vielfach nicht mit einer Veränderung der Standardabweichung einhergehen. Die Anstiege sowie Abfälle der mittleren Bewertung werden dementsprechend nicht von einzelnen Teilnehmern sondern vermutlich von der Mehrheit getragen. Allerdings bleibt zu berücksichtigen, dass der Mittelwert bzw. alle einbezogenen deskriptiven Maße zur Veranschaulichung eine sehr starke Reduktion darstellen. In den Plots aller einzelnen Bewertungsverläufe lässt sich maximal ein globaler Trend erkennen und es ist schwierig auch nur einzelne Rater-Paare zu finden, die ungefähr parallel ihre Bewertung verändern. Daher ist die positive Antwort auf die Forschungsfrage auf den globalen Bewertungsverlauf einzuschränken und nicht auf kurzfristige Veränderungen zu beziehen. An dieser Stelle wäre es wichtig, bei der Analyse wieder einen Schritt näher an die Originaldaten zu kommen und beispielsweise Gruppen ähnlich bewertender Probanden zu bilden.

5.3 Zusammenhang zwischen den kontinuierlichen und den retrospektiven Bewertungen

In diesem Kapitel werden die kontinuierlichen Performancebewertungen und retrospektiven Gesamturteile miteinander in Verbindung gebracht und es wird überprüft, in wie weit sich der Zusammenhang durch die in Abschnitt 2.6 dargestellten Modelle erklären lässt. Dazu wird zunächst definiert, welche Zeitpunkte der kontinuierlichen Bewertungen für diese Analyse als Bühnenauftritts-, Anfangs- und Endwertung sowie als Peak aufgefasst werden. Im Anschluss darauf aufbauend wird geprüft, welche Theorie bzw. welche Theorien zur Bildung eines Gesamturteils sich anhand

der erhobenen Daten bestätigen lassen und wie gut sie die abschließende Bewertung tatsächlich anhand der Zeitreihendaten vorhersagen können.

5.3.1 Methodische Vorüberlegungen

Für alle Zeitpunkte, denen einem oder mehreren Modellen zufolge eine größere Bedeutung für die Bildung eines abschließenden Urteils über die Performance zukommt, ergibt sich ein Problem, nämlich von welcher Latenzzeit der Zeitreihen ausgegangen werden soll. Selbst für die Theorie, dass die Zuhörer am Ende den Mittelwert aus allen Einzelwertungen bilden, ist die Latenzzeit relevant und die Grenzen des Zeitintervalls festzulegen. Als erster Eindruck kann einerseits die anfängliche Bewertung des Bühnenauftrittsverhaltens, andererseits aber auch die erste Bewertung im Sinne der ersten Bewertungsänderung ab dem ersten Ton gewertet werden. Allerdings ist diese Vorgehensweise nicht zielführend, da sich nicht bei allen Probanden anhand der Daten einzelne Bewertungsänderungen isolieren lassen. Beispielsweise funktioniert es nicht, wenn, während der erste Ton beginnt, über einen gewissen Zeitraum durchgehend die Bewertung immer weiter nach oben angeglichen wird oder innerhalb der ersten halben Minute nach dem ersten Ton überhaupt keine Veränderung vollzogen wird. Darüber hinaus ist es ungünstig einzelne Zeitpunkte für sich zu verwenden, da auch versehentliche kurzzeitige Veränderungen nicht ausgeschlossen werden können. Insofern muss an dieser Stelle von außen eine erwartete Latenz vorgegeben werden, da deren individuelle Bestimmung des Zeitpunkts für die erste Bewertung nicht möglich ist.

Die Bestimmung einer durchschnittlichen Latenzzeit anhand der Daten erweist sich als fast ebenso schwierig, wie die individuelle erste Bewertung in den Daten zu finden. Die Reaktion auf den ersten Ton jeder Performance ist inhaltlich die einzige Stelle, an der eine solche Berechnung sinnvoll funktionieren kann, da der Beginn des Bühnenauftritts zeitlich weit weniger genau festgelegt werden kann. Viele Reaktionen darauf passieren sehr zögerlich, oft erst deutlich später, aber andererseits starten auch einige Probanden vorzeitig mit der Bewertung. Sonstige Zeitpunkte innerhalb der Musikstücke können noch schlechter ausgesucht werden um eine Latenz zu berechnen, da dort keine Chance mehr besteht, eine Bewertungsänderung eindeutig als Reaktion einem bestimmten Ereignis zuzuordnen. Folglich wurde, wie oben bereits beschrieben, versucht für jede Person die erste Reaktion auf den Beginn des Stücks in den Daten zu finden bzw. die Zeit, die bis zu dieser Reaktion vergeht, wobei es nicht allzu problematisch ist, wenn dies in einigen Fällen nicht funktioniert. Allerdings gelang es nicht, eine sinnvolle Mindestgröße dieser Bewegung von

einem Sample zum nächsten festzulegen, die sowohl versehentliche minimale Bewegungen des Fingers nicht mitzählte, aber gleichzeitig auch keine aussagekräftigen Bewertungsänderungen übergang. Das passierte immer dann, wenn das Ausmaß der Änderung auf das einzelne Sample gesehen klein waren, da sie langsam passierten. Nach dieser Definition gab es immer Probanden, die sehr lange brauchten, oft über 30s, und für einige, bei denen der erste Ton mitten in einer Bewegung begann, kam als Latenzzeit Null heraus. Die übrigen Werte streuten sehr und es gab nie viele Werte, die nah am errechneten Mittelwert bzw. Median waren und diese beiden Größen variierten außerdem noch stark zwischen den Performances. Folglich sind das arithmetische Mittel und der Median in diesem Fall ungeeignet, eine typische Latenz anzugeben, die dann allgemein verwendet wird. Eine alternative Idee war, die Zeit von Beginn des Stücks bis zu der größten Veränderung innerhalb einer halben Sekunde, also von einem Sample zum darauf folgenden, innerhalb der ersten 20 s o. Ä. als Reaktionszeit zu interpretieren. Allerdings muss, auf die einzelne Person bezogen, die größte Bewegung nicht der ersten Reaktion entsprechen. Anhand der mittleren Bewertung diese Regel anzuwenden, ist auch nicht zielführend, da dieser stark von Ausreißern beeinflusst ist, gegenläufige Bewegungen einzelner Personen sich aufheben und es unmöglich ist, auszumachen wie viele einzelne Bewegungen für die Veränderung des Mittelwerts zu einem bestimmten Zeitpunkt verantwortlich sind. Bzgl. des Medians entfällt das Problem der Ausreißer, allerdings ist dieser im Verlauf immer stark abhängig von dem einen Bewertungsverlauf, der gerade den mittleren Rang innehat, und von Übergängen von einem Bewertungsverlauf zum anderen. Insofern sagt er noch weniger darüber aus, wie viel Bewegung zu einem bestimmten Zeitpunkt stattfindet. Letztlich lässt sich hier keine Lösung finden, zumal eine erste Entscheidung über die Musik auch sein kann, dass sie ebenso zu bewerten ist, wie zuletzt beispielsweise das Einstimmen oder das Ansetzen des Instrumentes, welches noch zum Bühnenauftrittsverhalten gehört. Infolgedessen wurde entschieden, von einer Latenzzeit von fünf Sekunden auszugehen, analog zu Schäfer et al. (2014), da dies die mittlere Einschätzung innerhalb der in diesem Kontext zitierten Literatur ist (vgl. Abschnitt 2.5.1) und da sich die weitere Auswertung insgesamt methodisch an dieser Studie orientiert. Diese Vorgehensweise ist nicht zufriedenstellend, da sie die Güte der Ergebnisse maßgeblich verschlechtern kann, aber in diesem Kontext nicht anders möglich, da belastbarere Werte für die Latenzzeit nicht vorliegen. Es müsste eine separate Studie durchgeführt werden, alleine um die Latenzzeit zu untersuchen, indem beispielsweise absichtlich sehr auffällige Fehler bzw. Aussetzer innerhalb einer Performance präsentiert würden,

sodass eine Zuordnung der Ereignisse einfacher wird. Gerade die anfängliche Latenzzeit könnte auch von Latenzzeiten im Stückverlauf abweichen, weil noch keine vorige Bewertung, an der sich die Probanden orientieren können, vorliegt.

Ein durch dieses äußere Kriterium gewählter einzelner Zeitpunkt, bei dem es Zufall ist, ob er auf den einzelnen Probanden bezogen den Zeitpunkt der ersten Bewertung darstellt, ist nicht verlässlich genug für die weitere Auswertung. Daher wird für den Zeitpunkt *Anfang* – und auch für die Endbewertung – jeweils ein Zehn-Sekunden-Fenster betrachtet. Bei der verwendeten Sample-Rate von 2 Hz entspricht dies 21 einzelnen Bewertungen, da die jeweiligen Randpunkte des Zeitintervalls mit berücksichtigt werden. Für die Auswertung wird jeweils die mittlere Bewertung in den angegebenen Zeiträumen betrachtet. Die genauen Definitionen der Zeitpunkte bzw. -fenster sind in Tabelle 5.4 aufgelistet und, um eine hohe Genauigkeit sicherzustellen, als Intervalle dargestellt. Bei der Festlegung der Zeitfenster, die in der Auswertung betrachtet werden sollen, ist die angenommene Latenzzeit folgendermaßen berücksichtigt: Es wird ausgeschlossen, dass zu Beginn des Stückes versehentlich Bewertungen zum reinen Bühnenverhalten der Musik zugewiesen werden, indem erst fünf Sekunden nach Beginn des Stückes die Bewertungen zur Variable *Anfang* hinzugerechnet werden. Am Ende des Stückes wird dies für den Bühnenabgang genauso gehandhabt. Auf der anderen Seite endet die Endbewertung tatsächlich mit dem letzten Ton bzw. der Aufgang mit dem ersten Ton, da die Personen auch schneller reagieren könnten. Dadurch geht möglicherweise die tatsächliche Bewertung des letzten Tons verloren und die Bewertungen von Personen, die mit einer deutlich größeren Latenz als fünf Sekunden reagieren, könnten falsch interpretiert werden. Dennoch scheint dies der beste Kompromiss zu sein und die größten Fehlerquellen auszuschließen. Würde man das Zeitfenster noch größer gestalten, würde man zu Beginn bei vielen Personen den wirklichen ersten Eindruck übergehen und erst zu spät einsteigen. Generell sind die Bewertungsänderungen, wie im Abschnitt 5.2 sowie bei dem Versuch, die Latenzzeit aus den Daten zu berechnen, dargestellt, eher von geringem Ausmaß und Häufigkeit, sodass größere Veränderungen der Ergebnisse bei leicht veränderter Festlegung der Zeitfenster unwahrscheinlich sind. Problematisch ist diese Definition allenfalls für den Abgang der Musiker, der dadurch erst fünf Sekunden nach dem letzten Ton betrachtet wird. Da dieser oft kurz war bzw. von den Zuhörern dessen Bewertung schnell abgebrochen wurde und dabei die bereits in Abschnitt 5.2 erwähnten, abrupt niedrigeren Bewertungen entstanden, die nicht ausgewertet werden sollten, bleibt oft nur ein sehr kurzes Zeitfenster. Allerdings ist dieses Problem nur wenig

5.3 Zusammenhang zwischen kontinuierlichen und retrospektiven Bewertungen

Zeitpunkte	0.00	Anfang des Videos, ungefähre Beginn des Aufgangs
	t_{anf}	Beginn des ersten Tons
	t_{end}	Ende des letzten Tons
Bewertungen	Aufgang	Mittelwert aller Wertungen für $t \in [5 s, t_{anf}]$
	Anfang	Mittelwert aller Wertungen für $t \in [t_{anf} + 5 s, t_{anf} + 15 s]$
	langer Anfang	Mittelwert aller Wertungen für $t \in [t_{anf} + 5 s, t_{anf} + 30 s]$
	Ende	Mittelwert aller Wertungen für $t \in [t_{end} - 5 s, t_{end}]$
	Abgang	Mittelwert aller Wertungen für $t \geq t_{end} + 5$
	Mittelwert Spielzeit	Mittelwert aller Wertungen für $t \in [t_{anf} + 5 s, t_{end}]$
	Peak	Das Maximum aller Bewertungen innerhalb der Spielzeit mit Ausnahme der ersten 5 s, falls dessen Absolutbetrag größer ist, als der des Minimums ebendieser Bewertungen, sonst das Minimum.
	Peak-End	Mittelwert aus Peak und Ende

Tabelle 5.4: Definitionen der betrachteten Einflussfaktoren

relevant, da die Bewertung des Abgangs für die Überprüfung der Hypothesen keine Rolle spielt, sofern man davon ausgeht, dass als letzter Eindruck die letzte Phrase ausschlaggebender ist als das Bühnenverhalten anschließend. Aus den genannten inhaltlichen Gründen, kann der Abgang in dieser Form nicht in der Auswertung berücksichtigt werden. Da sich später in der Analyse eine geringe Relevanz des zehnssekündigen Anfangs ergab, wurde zusätzlich ein längeres Zeitfenster für den Anfang bis 30 s nach dem ersten Ton mit berücksichtigt, um zu Prüfen, ob der Anfang wirklich weniger ausschlaggebend war oder lediglich das Intervall zu kurz gewählt. Diese Auslegung der Latenz ist auch auf das Zeitfenster für die mittlere Bewertung innerhalb der Spielzeit sowie des Peaks übertragen worden. Die Definition des Peaks für eine bipolare Skala ist, nicht eindeutig und wurde, wie in Abschnitt 2.6.1 dargestellt, bislang umgangen. Für die Auswertung hier wurde er festgelegt als entweder die niedrigste oder die höchste Bewertung, und zwar diejenige, deren Absolutbetrag höher war, die also stärker von der neutralen Bewertung abwich. Die Peak-End-Wertung entspricht dem Mittelwert aus der Peak-Wertung und der Endwertung.

5.3.2 Vergleich der Modelle

Um die Hypothesen bzgl. der Theorien zum Zusammenhang der kontinuierlichen und retrospektiven Bewertungen zu prüfen, wurde jeweils die Pearson-Korrelation

zwischen der Gesamtbewertung und der jeweiligen Bewertung eines Abschnitts bzw. der Peak- und der Peak-End-Wertung berechnet (s. Tabelle 5.5). Dabei wurden alle Stücke miteinbezogen, also ist jeder Rater in dieser Statistik fünfmal berücksichtigt. Die Bewertungen zu allen betrachteten Zeitpunkten bzw. -intervallen korrelieren hochsignifikant mit der Gesamtbewertung, wobei für Ende, Mittelwert der Spielzeit, Peak und Peak-End die Koeffizienten größer als .92 sind und somit jeder Faktor einzeln bereits mind. $r^2 = 85\%$ der Varianz der Gesamtbewertung erklärt. Am geringsten korreliert der Bühnenaufgang mit der Gesamtbewertung, um einiges stärker bereits der Anfang, allerdings dennoch nicht vergleichbar zu den bereits genannten Faktoren. Dass die Zusammenhänge insgesamt stark sind, ist nicht verwunderlich, da jeweils das gleiche bewertet wurde, wenn auch zu unterschiedlichen Zeitpunkten, und, da das Gesamturteil auf Basis der vorigen Bewertungen für die einzelnen Zeitpunkte getroffen wurde bzw. zumindest so getroffen werden konnte. Insofern ist die Korrelation von .407 zwischen dem Aufgang und der abschließenden Bewertung in diesem Kontext als schwacher Zusammenhang zu sehen und auch der Anfang ist im Vergleich zu den übrigen Faktoren wenig ausschlaggebend. Eventuell brauchten einige Probanden die 15 Sekunden ab Beginn des Stücks zur Meinungsbildung und gaben erst ab dann ihre Bewertung der Musik an, sodass diese in der Variable Anfang kaum oder gar nicht mehr berücksichtigt wird. Damit an dieser Stelle nicht die Hypothese zum Primacy Effekt fälschlicherweise verworfen wird, weil die zehn Sekunden für den Anfang ein zu kurzer Zeitraum waren, wurde zusätzlich noch der längere Anfang berücksichtigt, der immerhin etwas näher an die .9-er Korrelationen herankommt. Dennoch kann auch für diesen längeren Zeitraum nicht von einem gleichstarken Zusammenhang wie beispielsweise zwischen Ende und retrospektiver Wertung ausgegangen werden. Schwieriger wird es hingegen bei den vier höchsten Korrelationen zu entscheiden, ob man den Unterschieden der einzelnen Koeffizienten noch eine Bedeutung beimessen kann, oder diese so gering sind, dass sie als zufällig angesehen werden müssen. Peak und Ende separat betrachtet – mit einer Differenz der Korrelationskoeffizienten von 0.004 – sind anhand von dieser Analyse als gleichwertig anzunehmen und ebenso die Peak-End-Wertung, also der Mittelwert aus beiden, sowie der Mittelwert über die gesamte Spielzeit mit einer Differenz der Koeffizienten von 0.06. Insgesamt korreliert Peak-End knapp am stärksten von allen betrachteten Größen mit der Abschlusswertung. Allerdings ist schwer zu sagen, ob nun zwischen diesen beiden Gruppen an Faktoren noch ein inhaltlich bedeutsamer Unterschied besteht, oder ob sogar alle vier als eine Gruppe gleich wichtiger Faktoren angesehen werden sollten. Insgesamt sind die Korrelationen, selbst unter

5.3 Zusammenhang zwischen kontinuierlichen und retrospektiven Bewertungen

Korrelationen mit der Gesamtwertung	Aufgang	Anfang	langer Anfang	Ende	Mittelwert Spielzeit	Peak	Peak-End
Pearson-Korrelation	.407	.630	.704	.926	.944	.922	.950
Sig. (2-seitig)	≤.001	≤.001	≤.001	≤.001	≤.001	≤.001	≤.001
N	127	127	127	127	127	127	127

Tabelle 5.5: Korrelationen der Bewertungen zu einzelnen Zeitpunkten mit der Gesamtwertung

Berücksichtigung, dass die Variablen inhaltlich kausal zusammenhängen, sehr groß, also im Bezug auf relative Bewertungen die Übereinstimmungen sehr hoch.

Korrelationen geben hingegen nicht an, inwiefern die Bewertungen von den absoluten Skalenwerten her ebenfalls ähnlich sind, sodass diese hier zusätzlich betrachtet werden. Zunächst wurden dafür als neue Variablen die Differenz der jeweiligen Faktoren, beispielsweise der Endwertung, und der Gesamtwertung gebildet. Diese Differenz gibt bei einem positiven Wert an, um wie viel höher bzw. bei einem negativen Wert um wie viel niedriger der jeweilige Zeitpunkt im Vergleich zum abschließenden Rating bewertet wurde. Um nun vergleichen zu können, für welche Variable die Abweichung am größten bzw. geringsten ist, wurde der Mittelwert des Absolutbetrags gebildet, also die mittlere absolute Abweichung berechnet, anstatt die Mittelwerte der Variablen zu betrachten, damit sich positive und negative Abweichungen nicht ausgleichen. Das Balkendiagramm in Abb. 5.17, für das die Variablen nach mittlerer Abweichung absteigend sortiert wurden, zeigt ein ähnliches Verhältnis zwischen den einbezogenen Faktoren, wie zuvor die Korrelationen. Es sind deutliche Stufen zu erkennen, sodass wiederum der Aufgang die größte Abweichung aufweist, der Anfang, unabhängig von der Länge, etwas näher an der Abschlusswertung ist, aber eindeutig die übrigen vier Faktoren die beste Übereinstimmung haben. Wie bereits bei den Korrelationen ist die Peak-End-Bewertung knapp am nächsten an dem retrospektiven Rating. Der Mittelwert ist hier nicht der zweitbeste Indikator sondern der vierte nach End-Wertung und Peak-Wertung. Die Werte für die mittlere Abweichung reichen für diese vier Variablen von 12.6 für Peak-End bis 17.9 für den Mittelwert über die Spielzeit, was angesichts der 200-stufigen Skala nicht besonders viel ist. Hinzu kommt, dass die Skala für die Gesamtbewertung knapp halb so breit dargestellt wurde wie die Skala für die kontinuierliche Bewertung, die die volle Displaybreite einnahm, und bei letzterer entspricht eine Abweichung von 12.6 Punkten auf einem iPad mini lediglich

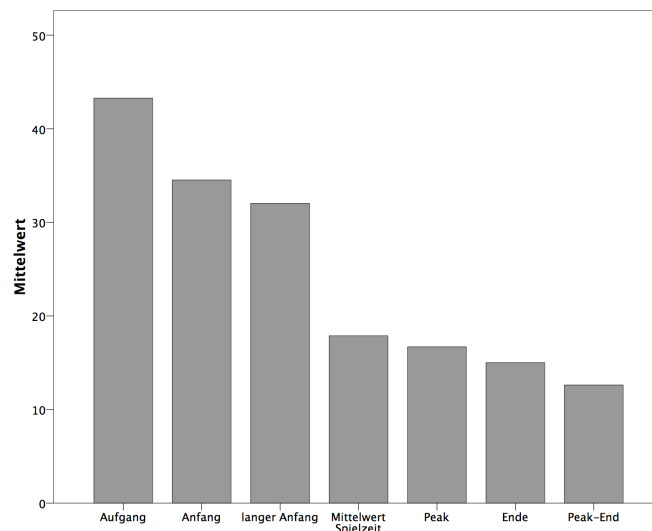


Abbildung 5.17: Mittlere absolute Abweichung von der Gesamtbewertung

einem Abstand von etwa 1.5 cm. Berücksichtigt man die dementsprechend noch geringeren Abstände, sowie das Umorientieren auf die schmalere Skala für die Gesamtbewertung, könnte eine Abweichung von weniger als 18 fast als Versuch, die gleiche Bewertung auf beiden Skalen zu treffen, interpretiert werden und entspricht jedenfalls nur einem geringen Unterschied.

In Abb. 5.18 sind als Boxplots die Verteilungen der Differenz-Variablen dargestellt, um einen Eindruck zur Richtung der Abweichungen zu bekommen. Dabei ist zu berücksichtigen, dass dadurch, dass alle Stücke miteinbezogen wurden, die Tendenzen eher zu verallgemeinern sind, es aber nicht möglich ist, Zusammenhänge mit den einzelnen Bewertungsverläufen zur Erklärung des Vorzeichens des Medians herzustellen. Abweichungen nach oben oder unten, die durch eher steigende oder eher fallende Bewertungsverläufe entstanden sind, beispielsweise am Anfang, werden sich bzgl. des Medians ausgleichen. Es bleibt darüber hinaus unberücksichtigt, inwiefern der Peak ein Minimum oder Maximum war, sodass sich Abweichungen in unterschiedliche Richtungen ebenfalls nur anhand der Streuung ablesen lassen, der Median dadurch aber neutraler wird. Es ist zu erkennen, dass der über alle Stücke verallgemeinerte Bewertungsverlauf leicht steigend über die gesamte Zeit gewesen sein muss, da Aufgang und Anfang am deutlichsten nach unten abweichen, auch wenn dies bei der großen Streuung keinesfalls auf einzelne Stücke übertragen werden kann. Die Endwertung entspricht im Median am genauesten der Abschlussbewertung und folglich ist der Mittelwert über die gesamte Spielzeit durch den schlechter

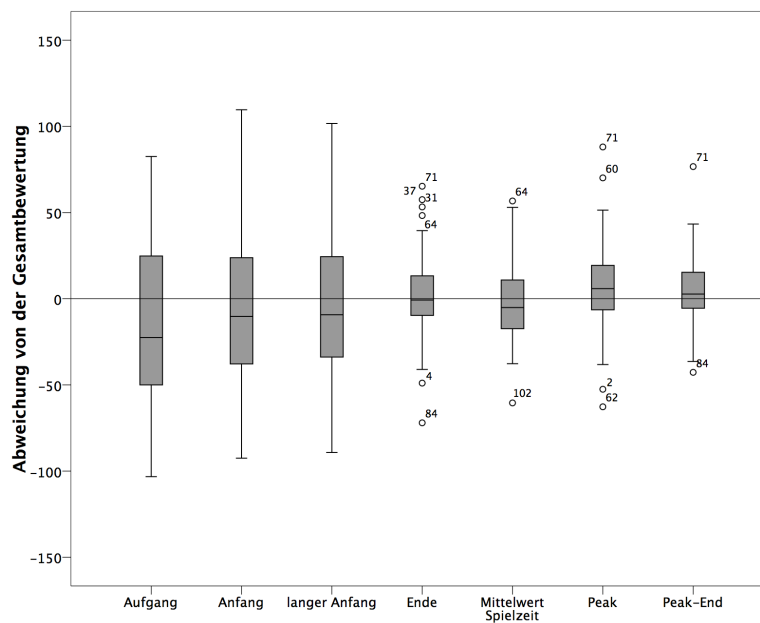


Abbildung 5.18: Differenzen der Bewertungen zu einzelnen Zeitpunkten und der Gesamtbewertung

bewerteten Anfang leicht unter der Gesamtbewertung ist. Die Peak-Bewertung ist demnach für die meisten Stücke und Personen ein Maximum, sodass der Median positiv ist. Die Peak-End-Bewertung liegt zwischen Peak- und End-Bewertung und ist im Median leicht positiv ist. Der Median ist hier betragsmäßig für fast alle Faktoren sehr klein, aber – wie bereits erläutert – sollten für die Abweichung die Beträge berücksichtigt oder alternativ die quadratische Abweichung berechnet werden, damit Abweichungen in verschiedene Richtungen sich nicht ausgleichen.

Um Aussagen bzgl. der Hypothesen zu treffen, müsste nun herausgestellt werden, welche Hypothese die beste ist. Durch den Vergleich der Korrelationskoeffizienten sowie der mittleren absoluten Abweichung ist bereits zu erkennen, dass ein Primacy Effekt, unabhängig davon, ob der längere oder kürzere Anfang oder sogar bereits den Bühnenaufgang berücksichtigt wird, nicht das beste Modell ist. Die Korrelationen zwischen den Bewertungen dieser Zeitintervalle und der Gesamtbewertung sind deutlich niedriger als die der anderen betrachteten Zeitpunkte mit dem abschließenden Urteil. Die mittleren absoluten Abweichungen sind im Vergleich deutlich höher und das Modell erklärt im besten Fall, für den langen Anfang, knapp 50% der Varianz der Gesamtbewertung ($r^2 = .496$), während die Varianzaufklärung

der übrigen Modelle jeweils mind. 85% beträgt. Somit wird die Hypothese 3c, die anfängliche Bewertung erkläre die retrospektive Bewertung am besten, verworfen.

Schwieriger wird es, unter den übrigen Modellen ein bestes zu finden bzw. zu entscheiden, ob sie gleichwertig sind oder es bedeutsame Unterschiede gibt. Es wäre an dieser Stelle denkbar, analog zu der Vorgehensweise von Schäfer et al. (2014), mit multipler linearer Regression mit der Gesamtbewertung als abhängige und den übrigen Zeitpunkten als unabhängige Variablen zu untersuchen, welche Zeitpunkte für die abschließende Bewertung am wichtigsten sind. Allerdings ist dies aufgrund von Multikollinearität nicht problemlos möglich. Abschnitt 5.3.2 zeigt, dass die End-Wertung, Peak-End-Wertung und durchschnittliche Wertung mit $r \geq .9$ untereinander sehr hoch korrelieren, sodass bereits per Definition Multikollinearität vorliegt (vgl. Pallant, 2005, 142f.).

Dass die verschiedenen Zeitintervalle aus der kontinuierlichen Daten stark miteinander korrelieren ist nicht weiter verwunderlich, da bei der Bildung des Mittelwerts über die gesamte Spielzeit auch der Anfang, das Ende und der Peak mitberücksichtigt werden bzw. Peak-End eine Linearkombination aus Peak- und Endbewertung ist. Außerdem sind bei kontinuierlich erfassten Bewertungen die aufeinander folgenden Datenpunkte automatisch miteinander korreliert, da sie als einzelne Positionen innerhalb einer Bewegung oder als identische Position, wenn keine Bewegung stattfand, nicht voneinander unabhängig sind. Ihr Abstand kann eine gewisse Größe nicht überschreiten und selten kommen überhaupt größere Abstände vor. Die Irrtumswahrscheinlichkeit darf daher nicht berücksichtigt werden, da die Voraussetzung der Unabhängigkeit der Messungen für den Signifikanztest durch das Echtzeitverfahren verletzt ist. Die Autokorrelationen zwischen aufeinander folgenden Samples führen auch bei allen weiteren Korrelationen der Zeitreihendaten untereinander zu einer Inflation der Korrelationskoeffizienten, weshalb diese alle vorsichtig zu interpretieren sind (genauere Darstellung der methodischen Probleme aufgrund von Autokorrelation s. Schubert, 2002; Upham, 2011). Beispielsweise wird die Endbewertung nur aus dem Mittelwert über die letzten zehn Sekunden der Stücke gebildet, allerdings werden ein gewisses Zeitintervall davor damit sehr stark zusammenhängen und dadurch wird der Zusammenhang der Endwertung mit der mittleren Wertung über die gesamte Spielzeit wiederum größer. Nicht nur die Daten aus den letzten zehn Sekunden werden doppelt ausgewertet, sondern auch das Zeitfenster direkt davor wird von beiden Variablen insofern berücksichtigt. Da die Korrelation zwischen der mittleren Bewertung des Anfangs bzw. der gesamten Spielzeit im Vergleich niedriger ist, obwohl für den langen Anfang 25 s berücksichtigt wurden, ist davon

auszugehen, dass die sehr hohen Korrelationen von $r \geq .9$ nicht ausschließlich auf Autokorrelationen zurückzuführen sind, sondern darüber hinaus ein inhaltlicher Zusammenhang vorhanden ist.

Das Wissen, dass die angewandte Methode der Datenerhebung eine gewisse Inflation der Korrelationskoeffizienten mit sich bringt, ändert allerdings nichts an der Problematik der Multikollinearität, die die Möglichkeiten der weiteren Datenanalyse mit Regressionsverfahren stark einschränkt. Bezieht man die vier Variablen langer Anfang, Ende, Peak-End und Mittelwert der Spielzeit in eine einzige multiple Regression zur Erklärung der Gesamtbewertung ein, so ist der Anfang der einzige Parameter, für den der Varianzinflationsfaktor mit $VIF = 2.9$ die üblicherweise angewandte Grenze von 10 nicht deutlich überschreitet. Für jeden anderen Faktor ist er größer als 14 ist und eine solche Analyse ist folglich nicht durchführbar. Reduziert man die Anzahl der berücksichtigten Variablen auf jeweils zwei, sind gemessen an dem Varianzinflationsfaktor bzw. auch an der Toleranz diejenigen möglich, die den Anfang und einen beliebigen anderen Parameter mit einbeziehen, und ergeben, dass der Anfang ein deutlich geringeres β -Gewicht hat. Allerdings war bereits klar, dass der Primacy-Effekt nicht das beste Modell für den Zusammenhang ist und insofern werden die Ergebnisse hier nicht im Detail dokumentiert. Für die Parameter, die tatsächlich von Interesse sind, sind bereits Regressionsmodelle mit nur zwei unabhängigen Variablen problematisch und nur in einem Fall (Ende und Mittelwert) ist der Varianzinflationsfaktor kleiner als 10. Wendet man aber die Definition von Multikollinearität als Korrelation der unabhängigen Variablen von $r \geq .9$ an (Pallant, 2005, 142f.), so wird sie zwar von dem Regressionsverfahren selbst nicht anhand der Kontrollparameter erkannt, ist aber dennoch vorhanden. Demnach ist auch bei einer Regression für Endwertung und Mittelwert über die Spielzeit, die mit $r(133) = .925$ korrelieren, die Voraussetzung als verletzt anzusehen. Auch wenn diese Modelle signifikant werden, ist eine Interpretation der Ergebnisse daher schwierig und die Aussagekraft stark eingeschränkt, sodass auf eine Auswertung verzichtet wird.

Es ist nicht ausgeschlossen, dass sich statistische Verfahren finden lassen, die an dieser Stelle noch eine Rangfolge der Variablen belegen können. Allerdings werden hier die Ergebnisse, dass eine Entscheidungsfindung mit Hilfe von Regressionsanalysen scheitert, so interpretiert, dass die Unterschiede zu klein sind, als dass es wichtig wäre, sie zu berücksichtigen. Es wurde bereits erläutert, dass es kausale Zusammenhänge zwischen den Variablen Ende, Peak-End und Mittelwert gibt, da teilweise die gleichen Daten ausgewertet werden, und, dass jede einzelne Variable

Pearson-Korrelationen	Ende	Peak-End	Mittelwert Spielzeit
langer Anfang	.656	.693	.784
Ende		.963	.925
Peak-End			.950

jeweils $p \leq .001$ und $N=135$

Tabelle 5.6: Interkorrelationen der unabhängigen Variablen

bereits einen großen Teil der Varianz der Gesamtbewertung erklärt. Dabei ist die Peak-End-Wertung knapp am besten mit einer Varianzaufklärung von $r^2 = .903$, gefolgt von der durchschnittlichen Bewertung ($r^2 = .891$) und mit etwas mehr Abstand, aber dennoch nah dran, die Endwertung ($r^2 = .857$). Danach könnte zusätzlich die Höhe des Peaks aufgeführt werden, die ebenfalls $r^2 = 85.0\%$ der Varianz der Gesamtbewertung erklärt, die allerdings bei den Theorien nie alleine genannt wird, sondern immer in Verbindung mit der Endwertung. Insofern werden die Hypothesen 3a, 3b und 3(d)i alle angenommen. Die Theorie, nach der der Mittelwert über die Zeitreihe ausschlaggebend für die Gesamtwertung ist, die Peak-End-Rule und bzgl. der Evolution der Bewertung die erste Hypothese werden somit jeweils als die Beste bestätigt. Die Hypothesen waren zwar so formuliert, um genau eine beste Theorie zu finden, es ist aber trotzdem logisch richtig, dass mehrere angenommen werden können, da „die beste Theorie“ im Sinne von „es gibt keine bessere Theorie“ zu verstehen ist. Damit ist gemeint, dass alle anderen Theorien schlechter oder gleich gut sein müssen, und genau dieser Fall ist hier – bis auf unbedeutende Abweichungen – eingetreten.

Offen bleibt nach dieser Analyse noch die zweite Hypothese der Evolution der Bewertung, also dass die Bewertungen spätestens nach 90s vollendet sind und sich nicht nennenswert weiter verändern. Dies ist mit den Ergebnissen aus Abschnitt 5.2.3 leicht zu widerlegen, da dort bereits gezeigt wurde, dass sich bei einigen Stücken die mittlere Bewertung bis zum Ende des Stückes weiterhin verändert. Es kann dennoch nicht unbedingt als Widerspruch zur Theorie ausgelegt werden, da der entscheidende Punkt vielmehr eine inhaltliche Frage ist. Ob die kontinuierlichen Bewertungen möglicherweise keine tatsächlich momentanen Wertungen sind, sondern akkumulierte Bewertungen, die zu jedem Zeitpunkt bereits die gesamte Performance bis dahin berücksichtigen, kann anhand der Daten weder widerlegt noch bestätigt werden. Nahegelegt wird die Vermutung einer akkumulierten Bewertung bzw. zunächst einmal einer abflachenden Bewertungskurve dadurch, dass die mittlere Bewertung über die gesamte Spielzeit viel stärker mit der Bewertung der

letzten zehn Sekunden korreliert als mit der ersten zehn oder sogar der ersten 25 s. Daraus folgt, dass entweder der Mittelwert schnell einen Wert, der der Schlusswertung ähnlich ist, erreichen muss, sodass die abweichende Bewertung am Anfang wenig ins Gewicht fällt, oder zwischendurch muss eine Abweichung der Bewertungen in die entgegengesetzte Richtung stattfinden. Wenn also die Bewertungen am Anfang unterhalb der durchschnittlichen Bewertung des gesamten Stücks sind, müssten sie zwischenzeitlich gegenüber dem Mittelwert auch nach oben abweichen um dies wieder auszugleichen. Die mittleren Verläufe der analysierten Stücke legen eher ersteres nahe, sodass der Anfang, speziell auch der Bühnenaufgang, von der mittleren kontinuierlichen Bewertung am stärksten abweicht. Dies ist aber aufgrund der im Bezug auf das gesamte Stück eher kurzen Dauer wenig ausschlaggebend für den Mittelwert. Da sich die Theorie der Evolution der Bewertung hier insofern nicht widerlegen lässt, muss sie entsprechend ihrer ersten, bestätigten Hypothese gleichwertig zur Peak-End-Rule und der Mittelwertannahme als ein bestes Modell angenommen werden.

5.3.3 Einfluss des Bühnenauftrittsverhaltens

Um zu untersuchen, ob es einen Zusammenhang zwischen der Bewertung des Bühnenauftrittsverhaltens und der Bewertung zu späteren Zeitpunkten sowie Gesamtbewertung gibt, wurden hierfür Pearson-Korrelationen berechnet (vgl. Tabelle 5.7). Dass die Wertung des Aufgangs zwar signifikant mit der Gesamtwertung korreliert, aber der Korrelationskoeffizient im Vergleich zu vielen anderen Einflussgrößen gering und insofern das Bühnenauftritts eher unwichtig für die Gesamtbewertung ist, wurde bereits im letzten Abschnitt gezeigt. Für das Ende der Performance ergibt sich eine geringfügig niedrigere Korrelation als für die Gesamtbewertung. Demzufolge ist auch der Einfluss auf diesen Zeitpunkt in der Performance als unwichtig zu betrachten. Relevanter hingegen könnten die Korrelationen der Bühnenauftrittswertung mit der anfänglichen Bewertung des Musikstücks sein, da diese deutlich größer sind, wenn auch nicht gleichauf mit den hohen Korrelationen über .9 zwischen Ende, Peak, Mittelwert über die Spielzeit und Gesamtbewertung.

Dabei darf für die Korrelationen zwischen Aufgang und den beiden Varianten des Stückanfangs aufgrund von Autokorrelation die Signifikanz wiederum nicht interpretiert werden und der Korrelationskoeffizient selbst nur dann, wenn er große Werte annimmt. Auch wenn zwischen dem Aufgang und dem Anfang des Stücks fünf Sekunden der kontinuierlichen Bewertung unberücksichtigt bleiben, können die Korrelationskoeffizienten dadurch erhöht sein. Für die Korrelation zwischen

5 Ergebnisse

Korrelationen mit dem Aufgang	Anfang	langer Anfang	Ende	Gesamtwertung
Pearson-Korrelation	.736	.674	.386	.407
Sig. (2-seitig)	≤.001	≤.001	≤.001	≤.001
N	135	135	135	127

Tabelle 5.7: Korrelationen bestimmter Zeitpunkte sowie der Gesamtbewertung mit dem Bühnenaufgang

den Bewertungen für den Bühnenaufgang und dem Ende des Musikstücks wurde dieses Problem nicht berücksichtigt, weil dazwischen mehrere Minuten vergehen, weshalb der Einfluss minimal sein sollte. Nun ist es schwierig, eine Entscheidung zu treffen, ob der Korrelationskoeffizient aufgrund von Autokorrelation für den kurzen Anfang am höchsten, sowie für den längeren Anfang noch immer erhöht ist, oder weil die Wahrnehmung des Bühnenauftritts die Einschätzung des Anfangs des Stücks beeinflusst. Allerdings ist es völlig problemlos möglich, innerhalb von 10 bzw. 25 s den Bewertungszeiger vom einen an das entgegengesetzte Ende der Skala zu bewegen. Da dies offensichtlich nicht passiert, kann der Zusammenhang nicht mehr vollständig auf Autokorrelation zurückgeführt werden, sondern muss auch eine inhaltliche Bedeutung haben. Insofern unterstützen die Ergebnisse die Hypothese 3e, dass die Bewertung von Bühnenauftrittsverhalten auch die Bewertung des Anfangs des Stückes beeinflusst, dieser aber abnimmt. Bereits zu dem längeren Anfang ist der Zusammenhang geringer und für die Gesamtbewertung unwichtig.

6 Zusammenfassung und Ausblick

Bei der Analyse der retrospektiven Bewertungen konnte übergreifend gezeigt werden, dass bei gleichzeitig großen absoluten Abweichungen eine hohe Konsistenz der Bewertungen vorliegt. Diese Aussage trifft sowohl auf die kleine Gruppe der Dozenten als auch auf das gesamte Publikum zu, sodass die Hypothesen 1a und 1b zur Inter-Rater-Reliabilität (s. Kapitel 3 auf Seite 45) bestätigt werden konnten. In beiden Fällen waren die Inter-Rater-Korrelationen im Mittel groß ($r = .847$ bzw. $r = .883$) und das Cronbach's α war sehr hoch ($\alpha = .945$ bzw. $\alpha = .990$). Auch wenn die Korrelationen der Selbsteinschätzungen der Musiker mit der Gesamtbewertung etwas niedriger waren als die Inter-Rater-Korrelationen innerhalb des Publikums, konnte hier eine mittlere bis hohe relative Übereinstimmung der Bewertungen bei z. T. großen absoluten Abweichungen bestätigt werden. Die Hypothese 1d, die Selbsteinschätzungen seien inkonsistent mit den Publikumsbewertungen, wurde daher verworfen. Darüber hinaus konnte gezeigt werden, dass die Musiker in der Lage sind einzuschätzen, in welche Richtung die Publikumsbewertung von ihrer eigenen abweichen wird. Trotz der hohen inneren Konsistenz der Publikumsbewertung ließen sich innerhalb des Publikums Gruppen mit unterschiedlicher Bewertungstendenz finden. Diese Gruppenunterschiede konnten nicht direkt für alle Stücke als signifikant gezeigt werden, sondern lediglich wenn Bewertungsunterschiede zwischen den Stücken herausgerechnet wurden. Die Bewertungsexpertise konnte als zusammenhängend mit einer negativen Bewertungstendenz belegt werden, wobei der Dozentenstatus, mit dem eine entsprechende Expertise einhergeht, ein besserer Indikator war als deren Selbsteinschätzung. Folglich wurde die Hypothese 1c, dass Probanden mit mehr Expertise kritischer bewerten, angenommen. Es ergab sich eine Effektstärke von $d = .68$ für den Mittelwertunterschied zwischen Dozenten und Studierenden, hingegen nur einer schwachen Korrelation der Selbsteinschätzung der Expertise mit den Gesamtbewertungen. Dass die Dozenten im Mittel schlechter bewerteten, könnte allerdings auch auf andere Ursachen zurückzuführen sein. Mit der gleichen Methode konnte darüber hinaus ein Geschlechterunterschied, dass Frauen besser bewerten, mit einer mittelgroßen Effektstärke belegt werden.

Die Analyse der kontinuierlichen Performancebewertungen konnte nicht alle diesbezüglichen Forschungsfragen beantworten. Es zeigte sich, dass die Schwankungen

der momentanen Standardabweichung über die Zeit gering sind, und sich daher kaum Stellen mit bedeutend größerer oder geringerer Streuung finden ließen. Die Reliabilität konnte aufgrund der angewandten Kriterien weder für einzelne Zeitpunkte noch global für alle Bewertungen eingeschätzt werden. Eine weitergehende Analyse wäre denkbar, ist im Rahmen dieser Arbeit allerdings nicht durchführbar. Der globale Bewertungsverlauf der einzelnen Zuhörer ist anhand der deskriptiven Statistiken betrachtet gleichförmig, beispielsweise ein kontinuierlicher Anstieg der ab einem gewissen Zeitpunkt stagniert. Allerdings lässt sich diese Aussage nicht auf kürzere Zeiträume und einzelne Bewertungsänderungen übertragen. Dabei bleibt zu berücksichtigen, dass die Daten bei der Betrachtung des mittleren Verlaufs stark reduziert wurden, und somit keinerlei Rückschlüsse auf einzelne Bewertungen möglich sind. Eine weitergehende Analyse sollte mehr in die Tiefe gehen und beispielsweise statt eines mittleren Verlaufs, Gruppen ähnlich bewertender Personen finden und nur gruppenweise mittlere Verläufe berechnen und vergleichen. Anhand der Videos der Performances ließen sich in einigen Fällen plausible Zusammenhänge zwischen den mittleren Bewertungen und den Auftritten finden. Speziell wenn Veränderungen über einen längeren Zeitraum passierten, war es meistens nicht möglich, diese anhand der Performances zu erklären. Insgesamt sind diese Ergebnisse nur ein Anfang der Datenauswertung und viele weitere Eigenschaften der kontinuierlichen Bewertungen könnten darüber hinaus untersucht werden.

Zwischen den kontinuierlichen und den retrospektiven Bewertungen wurden große Zusammenhänge festgestellt. Die Modelle Mittelwert, Peak-End-Rule und die Evolution der Bewertung erklärten jeweils mind. 85% der Varianz der Gesamtbewertung und es konnte statistisch nicht entschieden werden, welches das Beste ist, sodass alle drei als gleichwertig angenommen wurden. Lediglich die Hypothese 3c des Primacy-Effekts als beste Theorie konnte verworfen werden. Die zweite Hypothese bezüglich der Evolution der Bewertung, dass die Bewertung bereits nach 90s vollendet werde, wurde ebenfalls verworfen, allerdings ist damit die Theorie nicht grundsätzlich widerlegt. Die Frage, ob die mittlere Bewertung oder die Peak-End-Wertung ausschlaggebender für die Gesamtbewertung ist, kann sinnvoll anhand dieser Daten getestet werden, da sie die gleiche inhaltliche Bedeutung der kontinuierlichen Daten zugrunde legen. Sie basieren im Kontext Performancebewertung auf der Annahme, dass die kontinuierlichen Ratings sich auf die momentane Leistung der Musiker beziehen. Die Theorie der Evolution der Bewertung nach Thompson et al. (2007) geht von einem Akkumulationseffekt aus, sodass nicht nur die aktuelle Leistung bewertet, sondern auch die bisherige Leistung in der momentanen Bewertung integriert wird.

Demzufolge ist der Unterschied zwischen der Evolution der Bewertung und den übrigen Theorien vorwiegend ein inhaltlicher, sodass anhand des durchgeführten Experiments keine Aussage getroffen werden kann, welche Theorie die bessere ist. Es wäre sogar ein unterschiedlicher Umgang einzelner Probanden mit der Aufgabe der kontinuierlichen Ratings denkbar. Beispielsweise könnten einige Probanden tatsächlich die momentane Performance bewerten, dementsprechend stark auf Fehler reagieren, andere hingegen die gesamte Performance bis zum aktuellen Zeitpunkt im Blick haben, sodass Fehler von kurzer Dauer kaum ins Gewicht fallen. Es muss allerdings nicht als Nachteil gesehen werden, dass keine Entscheidung für die eine oder andere Theorie möglich ist. Die hohen Korrelationen zwischen dem Mittelwert, den einzelnen betrachteten Zeitpunkten, sowie der Gesamtbewertung sprechen für eine hohe intraindividuelle Reliabilität. Die einzelnen Probanden bewerten insofern zu den verschiedenen Zeitpunkten sehr konsistent. Darüber hinaus konnte die Hypothese 3e zum Einfluss des Bühnenauftritts bestätigt werden. Die Bewertung des Aufgangs hat zwar einen starken Zusammenhang mit der Bewertung des Anfangs des Stücks, allerdings nicht mit der Bewertung am Ende des Stückes oder der Gesamtbewertung.

Die Hypothese, die Platz (2014) aufstellte, dass ein als angemessen bewerteter Bühnenauftritt auch zu einer besseren Bewertung der gesamten Performance beitrage, wird somit von den vorliegenden Daten nicht unterstützt. Allerdings wurde auch nicht das Angemessenheitsurteil für den Aufgang erhoben, sondern eine Bewertung der Leistung und der Anfang wurde in keinem Fall negativ bewertet, sodass mögliche Unterschiede nicht zum Tragen kamen. Außerdem hatte das Publikum durch die kontinuierliche Bewertungsaufgabe einen zusätzlichen Grund, sich aktiv mit der Darbietung auseinanderzusetzen, sodass möglicherweise die Motivation zum Weiterhören aufgrund der Versuchsanweisung weniger relevant war. Bei diesem Versuchsdesign kann die These folglich auch nicht verworfen werden.

Abschließend ist festzuhalten, dass sowohl eine hohe innere Konsistenz der Publikumsbewertungen, die die Ergebnisse aus anderen Studien übertrifft (vgl. Davidson & Coimbra, 2001; Smith, 2004), als auch eine hohe intraindividuelle Konsistenz der Bewertungen vorliegt. Insgesamt bleibt aber bei allen Ergebnissen zu bedenken, dass die Stichprobe von $N_M = 27$ Zuhörern klein war und speziell die Anzahl der Stimuli mit nur fünf verschiedenen Performances sehr gering. Die Ergebnisse sind daher stark abhängig von den Darbietungen an dem Abend und möglicherweise wenig verallgemeinerbar. Die Arbeit kann insofern nur ein erster Schritt einer kontinuierlichen Untersuchung von Performancebewertungen sein.

Einerseits könnten anhand der vorliegenden Datensätze weitere Analysen durchgeführt werden. Beispielsweise könnte über Korrelationen zwischen den Zeitreihen oder der ersten Ableitung der Zeitreihen die Reliabilität bewertet werden. Obwohl beide Verfahren durch Autokorrelation und unterschiedliche Reaktionszeiten gewisse Schwächen haben, könnten damit Ergebnisse, die über die hier präsentierten hinausreichen, erzielt werden (vgl. Schubert, 2002; Upham, 2011). Speziell interessant wäre die Anwendung clusteranalytischer Verfahren auf die Echtzeit-Daten um Gruppen ähnlich wertender Probanden zu bilden und die weitere Analyse darauf aufbauend durchzuführen. Simple Verfahren sind allerdings kaum in der Lage, den kontinuierlichen Daten gerecht zu werden, und Optimierungen wären dementsprechend wichtig (vgl. Upham, 2011). Bei dem Zusammenhang der kontinuierlichen und retrospektiven Bewertungen könnte noch genauer untersucht werden, zu welchen Zeitpunkten die Peak-Bewertungen auftreten und stärker berücksichtigt, ob diese Maxima oder Minima sind. Möglicherweise könnten auch die einbezogenen Zeitfenster noch optimiert werden oder das von Rozin et al. (2004) vorgeschlagene, erweiterte Modell in die Analyse miteinbezogen werden. Dabei ist allerdings fraglich, ob so tatsächlich ein signifikant besseres Modell als die hier betrachteten gefunden werden kann.

Andererseits ergeben sich aus den Ergebnissen dieser Studie Implikationen für Folgestudien. Um weitergehend zu untersuchen, ob Frauen und Männer unterschiedlich bewerten und ab welchem Level die Expertise einen Einfluss auf die Bewertungen hat, wird ein Versuchsdesign mit deutlich mehr Probanden, etwa gleich großen Gruppen für die einzelnen Merkmale und mehr verschiedenen Stimuli benötigt. Auf eine kontinuierliche Bewertung der Performance könnte allerdings verzichtet werden.

Die entscheidende Frage, die geklärt werden muss, bevor kontinuierliche Verfahren zur Performancebewertung in anderen Kontexten eingesetzt werden, ist die, welche inhaltliche Bedeutung die momentanen Wertungen am ehesten haben. Sollen anhand der Echtzeitbewertungen Rückschlüsse auf die Performance gezogen werden, beispielsweise um die Musiker zu coachen, muss geklärt sein, ob sie tatsächlich die gewünschte momentane Bewertung angeben. Liegen akkumulierte Bewertungen vor, muss die Deutung der Bewertungsänderungen angepasst werden. Um diese Frage zu untersuchen, könnten Performances gefilmt und anschließend in Abschnitte zerlegt werden, die einzeln von Probanden bewertet werden. Anhand des Zusammenhangs zwischen den kontinuierlichen Bewertungen der gesamten Performance und den Bewertungen der einzelnen Abschnitte könnten Schlüsse darüber gezogen werden,

inwiefern die kontinuierlichen Daten akkumuliert sind. Wenn möglich, könnte die Performance auch in einer anderen Reihenfolge wieder zusammengesetzt und die jeweiligen kontinuierlichen Bewertungen beider Varianten miteinander verglichen werden. Zusätzlich könnte auch überprüft werden, ob kontinuierliche Bewertungen in der Live-Situation bzw. anhand der Video-Aufzeichnung eine hohe Übereinstimmung haben oder die Live-Situation für die Bewertung einen substantziellen Unterschied macht. Außerdem wäre es in der Live-Situation wichtig, mit einer Kontrollgruppe, die die Performances nur retrospektiv bewertet, zu untersuchen, ob die kontinuierliche Evaluation der Leistung einen Einfluss auf die Gesamtbewertung hat.

Literatur

- Bergee, M. J. (1993). A comparison of faculty, peer, and self-evaluation of applied brass jury performances. *Journal of research in music education*, 41(1), 19–27.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin, Heidelberg: Springer.
- Brittin, R. V. & Duke, R. A. (1997). Continuous versus summative evaluations of musical intensity: A comparison of two methods for measuring overall effect. *Journal of Research in Music Education*, 45(2), 245–258.
- Brittin, R. V. & Sheldon, D. A. (1995). Comparing continuous versus static measurements in music listeners' preferences. *Journal of Research in Music Education*, 43(1), 36–46.
- Davidson, J. W. (2009). Movement and collaboration. In *The Oxford handbook of music psychology*. (S. 364–376). Oxford library of psychology. Oxford University Press.
- Davidson, J. W. & Coimbra, D. D. C. (2001). Investigating performance evaluation by assessors of singers in a music college setting. *Musicae Scientiae*, 5(1), 33–53.
- Duke, R. A. & Colprit, E. J. (2001). Summarizing listener perceptions over time. *Journal of Research in Music Education*, 49(4), 330–342.
- Emmanuel, T. (2013). The Tall Fiddler. Zugriff 14. Juni 2015, unter <https://www.youtube.com/watch?v=nW9rmaGaaG4>
- Fredrickson, B. L. & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, 65(1), 45–55.
- Geringer, J. M. (1995). Continuous loudness judgments of dynamics in recorded music excerpts. *Journal of Research in Music Education*, 43(1), 22–35.
- Griffiths, N. K. (2008). The effects of concert dress and physical appearance on perceptions of female solo performers. *Musicae scientiæ: The journal of the European Society for the Cognitive Sciences of Music*, 12(2), 273–290.
- Häcker, H. (2013). Konsistenz des Verhaltens. In M. A. Wirtz (Hrsg.), *Dorsch. Lexikon der Psychologie* (S. 918). Verlag Hans Huber.

- Huang, J. & Krumhansl, C. L. (2011). What does seeing the performer add? It depends on musical style, amount of stage behavior, and audience expertise. *Musicae scientiæ: The journal of the European Society for the Cognitive Sciences of Music*, 15(3), 343–364.
- Juslin, P. N. & Timmers, R. (2010). Expression and communication of emotion in music performance. In *Handbook of music and emotion: Theory, research, applications*. (S. 453–489.). Oxford University Press.
- Kahneman, D., Fredrickson, B. L., Schreiber, C. A. & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, 4(6), 401–405.
- Kalies, C., Lehmann, A. C. & Kopiez, R. (2008). Musikleben und Live-Musik. In H. Bruhn, R. Kopiez & A. C. Lehmann (Hrsg.), *Musikpsychologie: Das neue Handbuch*. (S. 293–315). Rowohlt.
- Kluge, F. (Hrsg.). (2011). *Etymologisches Wörterbuch der deutschen Sprache* (25. Aufl.). Berlin: de Gruyter.
- Kopiez, R. (2005). Experimentelle Interpretationsforschung. In H. d. l. Motte-Haber & G. Rötter (Hrsg.), *Musikpsychologie. Handbuch der systematischen Musikwissenschaft, No. 3* (S. 459–514). Laaber-Verlag.
- Kopiez, R. (2008). Reproduktion und Interpretation. In H. Bruhn, R. Kopiez & A. C. Lehmann (Hrsg.), *Musikpsychologie: Das neue Handbuch*. (S. 316–337). Rowohlt.
- Kopiez, R. (2010). Performanceforschung. In H. d. l. Motte-Haber, H. v. Loesch, G. Rötter & C. Utz (Hrsg.), *Lexikon der systematischen Musikwissenschaft: Musikästhetik, Musiktheorie, Musikpsychologie, Musiksoziologie*. (S. 367–370). Laaber-Verlag.
- Kopiez, R., Dressel, J., Lehmann, M. & Platz, F. (2011). *Vom Sentographen zur Gänsehautkamera: Entwicklungsgeschichte und Systematik elektronischer Interfaces in der Musikpsychologie*. Tectum-Verl.
- Lehmann, A. C. (2014). Using admission assessments to predict final grades in a college music program. *Journal of research in music education*, 62(3), 245–258.
- Lehmann, M. & Kopiez, R. (2011). Der Einfluss der Bühnenshow auf die Bewertung der Performanz von Rockgitarristen. In R. F. Nohr & H. Schwaab (Hrsg.), *Metal Matters: Heavy Metal als Kultur und Welt* (S. 195–206). Münster: Lit Verlag.



- Louven, C. & Scholle, C. (in Vorbereitung). *emoTouch* für iPad: Ein flexibles, mobiles Forschungswerkzeug zur Erhebung kontinuierlicher Probandenratings in ein und zwei Dimensionen. *Musikpsychologie: Jahrbuch der Deutschen Gesellschaft für Musikpsychologie*, 25.
- Madsen, C. K. (1998 Winter). Emotion versus Tension in Haydn's Symphony No. 104 as Measured by the Two-Dimensional Continuous Response Digital Interface. *Journal of Research in Music Education*, 46(4), 546–554.
- McPherson, G. E. & Schubert, E. (2004). Measuring performance enhancement in music. In A. Williamon (Hrsg.), *Musical excellence: Strategies and techniques to enhance performance*. (Kap. 4, S. 61–82). Oxford University Press.
- Mikula, G. (2013). Konsistenz, innere. In M. A. Wirtz (Hrsg.), *Dorsch Lexikon der Psychologie* (S. 918). Verlag Hans Huber.
- Nagel, F. (2007). *Psychoacoustical and Psychophysiological Correlates of the Emotional Impact and the Perception of Music* (Diss., Hannover University of Music und Drama).
- Neuhoff, H. (2007). Die Konzertpublika der deutschen Gegenwartskultur: empirische Publikumsforschung in der Musiksoziologie. In H. de la Motte-Haber & H. Neuhoff (Hrsg.), *Musiksoziologie*. (S. 473–509). Handbuch der systematischen Musikwissenschaft. Laaber-Verlag.
- Pallant, J. (2005). *SPSS survival manual : a step by step guide to data analysis using SPSS for Windows* (12. Aufl.). Open University Press.
- Peddell, L. T. (2004). *Influence of conductor behavior on listeners' perception of expressiveness* (Diss., University of Minnesota).
- Peddell, L. T. (2008). Factors influencing listeners' perception of expressiveness for a conducted performance. *Bulletin of the Council for Research in Music Education*, (178), 47–61.
- Platz, F. (2014). *Wenn der Musiker erscheint : Der audiovisuelle Eindruck im Konzert*. Tectum.
- Platz, F. & Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music perception: An interdisciplinary journal*, 30(1), 71–83.
- Platz, F. & Kopiez, R. (2013). When the first impression counts: Music performers, audience and the evaluation of stage entrance behaviour. *Musica scientiæ: The journal of the European Society for the Cognitive Sciences of Music*, 17(2), 167–197.


- Redelmeier, D. A. & Kahneman, D. (1996). Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, *66*(1), 3–8.
- Rozin, A., Rozin, P. & Goldberg, E. (2004). The feeling of music past: How listeners remember musical affect. *Music Perception*, *22*(1), 15–39.
- Ryan, C., Wapnick, J., Lacaille, N. & Darrow, A.-A. (2006). The effects of various physical characteristics of high-level performers on adjudicators' performance ratings. *Psychology of music*, *34*(4), 559–572.
- Schäfer, T., Zimmermann, D. & Sedlmeier, P. (2014). How We Remember the Emotional Intensity of Past Musical Experiences. *Frontiers in Psychology*, *5*(911).
- Schlemmer, M. (2005). Audiovisuelle Wahrnehmung: Die Konkurrenz und Ergänzungssituation von Auge und Ohr bei zeitlicher und räumlicher Wahrnehmung. In H. de la Motte-Haber & G. Rötter (Hrsg.), *Musikpsychologie*. (S. 173–184). Handbuch der systematischen Musikwissenschaft. Laaber-Verlag.
- Schubert, E. (2001). Continuous measurement of self-report emotional response to music. In P. N. Juslin & J. A. Sloboda (Hrsg.), *Music and emotion: Theory and research* (S. 393–414). Oxford University Press.
- Schubert, E. (2002). Correlation analysis of continuous emotional response to music: Correcting for the effects of serial correlation. *Musicae scientiæ: The journal of the European Society for the Cognitive Sciences of Music*, 213–236.
- Schubert, E. (2010). Continuous self-report methods. In *Handbook of music and emotion: Theory, research, applications* (S. 223–253). Series in affective science. Oxford University Press.
- Schubert, E. (2013). Reliability issues regarding the beginning, middle and end of continuous emotion ratings to music. *Psychology of music*, *41*(3), 350–371.
- Sedlmeier, P. & Renkewitz, F. (2013). *Forschungsmethoden und Statistik. Ein Lehrbuch für Psychologen und Sozialwissenschaftler* (2., aktualisierte und erweiterte Auflage). Pearson.
- Smith, B. P. (2004). Five judges' evaluation of audiotaped string performance in international competition. *Bulletin of the Council for Research in Music Education*, (160), 61–69.
- Thompson, S. & Williamon, A. (2003). Evaluating Evaluation: Musical Performance Assessment as a Research Tool. *Music Perception*, *21*(1), 21–41.

- Thompson, S., Williamon, A. & Valentine, E. R. (2007). Time-dependent characteristics of performance evaluation. *Music perception: An interdisciplinary journal*, 25(1), 13–29.
- Upham, F. (2011). *Quantifying the temporal dynamics of music listening: A critical investigation of analysis techniques for collections of continuous responses to music* (Masterarbeit, McGill University).
- Wapnick, J., Campbell, L., Siddell-Strebel, J. & Darrow, A.-A. (2009). Effects of non-musical attributes and excerpt duration on ratings of high-level piano performances. *Musicae scientiæ: The journal of the European Society for the Cognitive Sciences of Music*, 13(1), 35–54.
- Wapnick, J., Darrow, A.-A., Kovacs, J. & Dalrymple, L. (1997). Effects of physical attractiveness on evaluation of vocal performance. *Journal of research in music education*, 45(3), 470–479.
- Wapnick, J., Darrow, A.-A. & Mazza, J. K. (1998). Effects of performer attractiveness, stage behavior, and dress on violin performance evaluation. *Journal of research in music education*, 46(4), 510–521.
- Wapnick, J., Jasinskas, L., Flowers, P. J. & Alegant, M. (1993). Consistency in piano performance evaluation. *Journal of research in music education*, 41(4), 282–292.
- Wapnick, J., Mazza, J. K. & Darrow, A.-A. (2000). Effects of performer attractiveness, stage behavior, and dress on evaluation of children’s piano performances. *Journal of research in music education*, 48(4), 323–335.
- Wapnick, J., Ryan, C., Campbell, L., Deek, P., Lemire, R. & Darrow, A.-A. (2005). Effects of excerpt tempo and duration on musicians’ ratings of high-level piano performances. *Journal of research in music education*, 53(2), 162–176.
- Windsor, W. L. (2009). Measurement and models of performance. In *The Oxford handbook of music psychology* (S. 323–331). Oxford library of psychology. Oxford University Press.
- Wirtz, M. A. (Hrsg.). (2014). *Dorsch - Lexikon der Psychologie*. (17. Aufl.). Huber.
- Wolf, A. & Kopiez, R. (2014). Do grades reflect the development of excellence in music students? The prognostic validity of entrance exams at universities of music. *Musicae scientiæ: The journal of the European Society for the Cognitive Sciences of Music*, 18(2), 232–248.
- Ybarra, O. (2001). When first impressions don’t last: The role of isolation and adaptation processes in the revision of evaluative impressions. *Social Cognition*, 19(5), 491–520.

Anhang

Fragebogen für das Publikum

	1. Die Bewertungen in diesem Fragebogen beziehen sich auf Beitrag Nr....							
	Test 1	Test 2	1	2	3	4	5	6
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	2. Bitte gib eine Gesamtwertung für diese Performance ab!							
	sehr schlecht							sehr gut
3. Bitte bewerte folgende Aussagen								
	stimme gar nicht zu				stimme voll zu			
Ich bin mit dem / einem der Musiker gut befreundet.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich mag das Stück.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich kenne das Stück gut.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Das Stück hat einen hohen Schwierigkeitsgrad.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Weiter								
Carolin Scholle, Universität Osnabrück – 2015								

	<p>4. War dies der letzte Musiker / das letzte Ensemble, das bewertet wird?</p> <p><input type="text" value="Nein, es kommt noch jemand!"/></p> <p><input type="text" value="Ja, der Bewertungsteil ist vorbei."/></p> <p>Carolin Scholle, Universität Osnabrück – 2015</p>
---	---

	<p>Vielen Dank für die Wertung!</p> <p>Die Antworten wurden gespeichert, du kannst jetzt das Fenster mit dem Knopf oben rechts schließen.</p> <p>Carolin Scholle, Universität Osnabrück – 2015</p>
--	--



Das iPad kann jetzt bis zum Ende des Vorspiels zugeklappt werden.

Das Vorspiel ist inzwischen bereits vorbei? Dann jetzt auf weiter klicken!

Weiter

Carolin Scholle, Universität Osnabrück – 2015



5. Wie alt bist Du?

Ich bin Jahre alt.

6. Welches Geschlecht hast du?

- weiblich
 männlich

7. Ich bin...

- Musikstudierende/r im Bachelor (1.-2. Semester)
 Musikstudierende/r im Bachelor (ab 3. Semester)
 Musikstudierende/r im Master of Education
 Musikstudierende/r im Master Musikwissenschaft
 Absolvent/in des Master of Education und nicht mehr an der Uni
 Instrumentallehrer/in
 künstlerische/r oder wissenschaftliche/r Mitarbeiter/in
 Professor/in
 Laienmusiker/in
 sonstiges:

Weiter

Carolin Scholle, Universität Osnabrück – 2015



8. An wie vielen vergleichbaren Vorspielsituationen hast du im vergangenen Jahr als Zuhörer teilgenommen?

Es zählen auch die, in denen du gespielt, aber auch den anderen zugehört hast, mit.

- an keiner
- 1-5
- 5-10
- mehr als 10

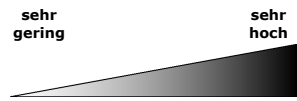
9. In wie vielen vergleichbaren Vorspielsituationen hast du im letzten Jahr selbst gespielt/gesungen?

- in keiner
- 1-5
- 6-10
- mehr als 10

Weiter

Carolin Scholle, Universität Osnabrück – 2015

10. Meine Erfahrung mit...



...der Bewertung von Performances schätze ich ein als:

...Instrumentalunterricht (als Lehrender) schätze ich ein als:

11. Bitte bewerte folgende Aussagen!

stimme gar nicht zu **stimme voll zu**

Es war für mich einfach, mit der App umzugehen.

Die App hat mich stark von den Performances abgelenkt.

Die kontinuierliche Bewertung hat mir geholfen, die Abschlusswertung zu geben.

Meine Nachbarn haben mich bei der Bewertung beeinflusst.

Ich benutze regelmäßig ein Smartphone oder Tablet.

Weiter



12. Du bist jetzt fast fertig. Wenn du noch Anmerkungen zu dem Experiment, dem Fragebogen oder der App hast, ist hier Platz dafür!

Weiter

Carolin Scholle, Universität Osnabrück – 2015



Du hast das Ende des Fragebogens erreicht. Danke für die Teilnahme!

Die Antworten wurden gespeichert, du kannst jetzt das Fenster mit dem Knopf **oben rechts schließen** und das iPad **abgeben**.

Carolin Scholle, Universität Osnabrück – 2015

Fragebogen für die Musiker

PUBLIKUMSBEWERTUNGEN

Musiker-Fragebogen

Korrekturen können, falls nötig, vorgenommen werden, indem das Kästchen ganz ausgemalt wird und ein neues Kreuz in das gewünschte Feld gesetzt wird.

Ich habe mitgespielt bei Beitrag Nr.	
Wie alt bist du?	Jahre
Geschlecht:	<input type="checkbox"/> männlich <input type="checkbox"/> weiblich
In welchem Studiengang bist du aktuell eingeschrieben?	<input type="checkbox"/> Bachelor <input type="checkbox"/> Master
In welchem Semester bist du insgesamt?	
Hast du heute auf deinem Hauptinstrument vorgespielt?	<input type="checkbox"/> ja <input type="checkbox"/> nein
Bereitest du dich mit diesem Stück aktuell auf eine Prüfung oder einen Wettbewerb vor?	<input type="checkbox"/> ja <input type="checkbox"/> nein

	stimme gar nicht zu					stimme voll zu				
Ich habe das Stück gut vorbereitet.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Ich habe mich wegen der Studie besser vorbereitet als sonst.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Das Stück ist für mich sehr anspruchsvoll.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Ich konnte die Wertungssituation mit den iPads während des Auftritts überwiegend ignorieren.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Der Anblick der iPads hat mich nervös gemacht.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Ich war heute nervöser als meistens bei den Vorspielen.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Ich habe heute mehr auf mein Verhalten auf der Bühne geachtet (Aufgang, Bewegung beim Spielen,...).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Die Vorspielsituation war für mich wenig anders als sonst bei den Vorspielen.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Ich bin zufrieden damit, wie ich heute gespielt habe.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

	viel schlechter	etwas schlechter	genauso gut	etwas besser	viel besser
Im Vergleich zu den Proben lief das Stück heute...	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Gib eine Wertung ab, indem du an der entsprechenden Stelle einen senkrechten Strich durch den Balken zeichnest!

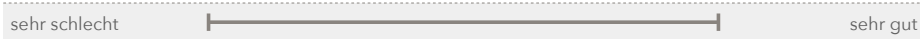
Beispielskala, wie eine Wertung aussehen soll:



Ich würde die Performance unseres Ensembles (bzw. meine Solo-Performance) so bewerten:



Ich denke, das Publikum hat die Performance unseres Ensembles (bzw. meine Solo-Performance) so bewertet:



Wenn du noch Anmerkungen zum Experiment oder dem Fragebogen hast, ist hier Platz dafür!

Vielen Dank für die Teilnahme!

Ich versichere, dass ich die eingereichte Master-Arbeit selbstständig und ohne unerlaubte Hilfe verfasst habe. Anderer als der von mir angegebenen Hilfsmittel und Schriften habe ich mich nicht bedient. Alle wörtlich oder sinngemäß den Schriften anderer Autoren entnommenen Stellen habe ich kenntlich gemacht.

Osnabrück, 21. September 2015